DAWN-CAST: DYNAMICAL ADAPTIVE WAVELET NETWORK FOR PRECIPITATION NOWCASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Precipitation nowcasting, the prediction of rain over the next minutes to hours, is critical for warning of flash floods and severe convective storms. Typically, weather radar data is used for this task and modeled as a video prediction task. Within an individual storm, rainfall patterns do not evolve uniformly: coherent, large-scale structures drift smoothly, whereas small-scale convective variations change suddenly and lose correlation within just a few frames. Existing nowcasting models capture multiscale spatial structure through Fourier decomposition, wavelets, convolutions, or generative modeling, yet they do not assign each precipitation scale its own temporal pathway. We propose DAWN-Cast, a **D**ynamical **A**daptive **W**avelet **N**etwork that lets every scale learn its own temporal evolution. Its Wavelet Guided Temporal Modeling (WGTM) block splits radar features into wavelet scale families and assigns each its own Frequency Adaptive Temporal (FAT) block, with no interaction between scales before reconstruction. Each FAT block pairs a nonlinear MLP with an adaptive, history-selective Gabor mapping. A refinement stage then reconciles the independently evolved scales by combining global spectral and local spatial interactions. Across four benchmark datasets, DAWN-Cast achieves the best result on 19 of 20 reported metrics, and ablations show that per-scale temporal operators outperform a shared one. DAWN-Cast also generalizes zero-shot to unseen radar domains and adapts to a new dataset by updating only 0.71% of its parameters.

1 INTRODUCTION

Precipitation nowcasting refers to forecasting rainfall over the coming minutes to a few hours. This is usually done using weather radar observations and is often framed as a video prediction problem. Reliable nowcasts support flood preparedness, aviation safety, and traffic and energy management Veillette et al. (2020); Shi et al. (2015). The task remains hard: the atmosphere evolves rapidly, precipitation structures within a single storm evolve at different paces, and the underlying dynamics vary from one climatic region to another Sun et al. (2014).

Deep learning has become the dominant approach (Sec. 2), and the multiscale structure of radar fields has been exploited extensively in the spatial domain through Fourier decomposition, wavelet transforms Kumar & Foufoula-Georgiou (1994), convolutions, and generative models. Recent decomposition-based nowcasting models go further and give separate networks to separate components: AlphaPre splits the amplitude from phase Lin et al. (2025), and WADEPre splits the approximation from the details of a wavelet decomposition Liu et al. (2026). However, AlphaPre's split is not by spatial scale, and WADEPre routes all fine scales through a single shared branch that couples them, so the temporal evolution of each scale is never modeled on its own terms. This matters because coherent large-scale structures drift smoothly, whereas fine-scale convective variability changes abruptly (Appendix E quantifies this on SEVIR).

A second gap concerns transfer across regions. A model trained on one radar dataset typically needs substantial retraining before it works in another climate, and whether organizing temporal dynamics by scale eases this transfer has not been examined.

We address both gaps with DAWN-Cast, a Dynamical Adaptive Wavelet Network whose temporal modeling is organized by scale. A wavelet decomposition splits lifted radar features into scale

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

families (Appendix D), each family is forecast by its own temporal operator, and a refinement stage reconciles the scales after reconstruction.

What is new. Wavelet decomposition, Gabor mappings, and spectral refinement have each been studied before, and WADEPre already decomposes precipitation into wavelet bands Liu et al. (2026). Our novelty is not decomposing precipitation into scales, but assigning each scale its own temporal dynamics: every wavelet scale family receives an independent temporal operator and remains isolated from the other families until reconstruction. This design follows from the scale-dependent temporal persistence of precipitation (Appendix E) and is tested through shared-versus-separate temporal ablations (Sec. 5.1) and through cross-domain transfer against a parameter-matched shared-operator variant (Sec. 5.3.1).

We make three contributions:

1. We introduce **scale-specific temporal modeling** for precipitation nowcasting, implemented in the Wavelet Guided Temporal Modeling (WGTM) block, whose per-family Frequency Adaptive Temporal (FAT) operators combine a nonlinear MLP with a history-selective Gabor mapping, followed by a spectral–local refinement stage.
2. DAWN-Cast outperforms several baselines in SEVIR, MeteoNet, Shanghai, and CIKM radar datasets. And ablations show that separate per-family temporal operators outperform both a shared operator and per-family MLPs.
3. DAWN-Cast transfers zero-shot to unseen radar domains (e.g., trained on SEVIR, tested on Shanghai) better than all baselines and a parameter-matched shared-operator variant, and adapts to a new domain by updating only 0.71% of its parameters (Sec. 5.3).

2 Related Work

Deep nowcasting models differ mainly in how they carry a radar history forward in time: recurrent models propagate a hidden state frame by frame Shi et al. (2015; 2017); Chang et al. (2021); Guen & Thome (2020), recurrence-free models predict all future frames in one pass using convolutions, attention, or Fourier mixing Gao et al. (2022a;b); Song et al. (2026); Wu et al. (2024), and generative models sample the fine-scale detail that deterministic training tends to average away Yu et al. (2024); Gao et al. (2023). Decomposition-based models give separate components their own networks. Specifically, AlphaPre for instance, separates Fourier amplitude from phase, but not spatial scales Lin et al. (2025).

Difference from wavelet-based nowcasting. The closest prior approach, WADEPre Liu et al. (2026), also decomposes the radar field with a wavelet transform and gives the approximation and detail bands separate networks. DAWN-Cast differs in two ways. First, WADEPre processes all detail levels through one shared temporal pathway and exchanges features between them, whereas DAWN-Cast assigns an independent temporal operator to each scale family and prevents cross-scale interaction until reconstruction. Second, DAWN-Cast's temporal operators have no spatial receptive field, whereas both WADEPre networks mix space; spatial interaction is deferred to the refinement stage. Table 10 summarizes these differences, and Appendix G compares the two models component by component.

3 Methodology

DAWN-Cast (Fig. 1) maps each input radar frame to c channels with a per-frame convolutional lifting $\mathcal{L}$, forecasts in this lifted space with the WGTM block (Sec. 3.2.1) and the refinement stage (Sec. 3.2.2), and maps the result back to radar frames with a projection $\mathcal{P}$.

3.1 Motivation

3.1.1 Why wavelet

Precipitation fields combine broad systems with localized, intermittent structures, and wavelets have long been used to separate their large- and small-scale fluctuations while retaining spatial localiza-

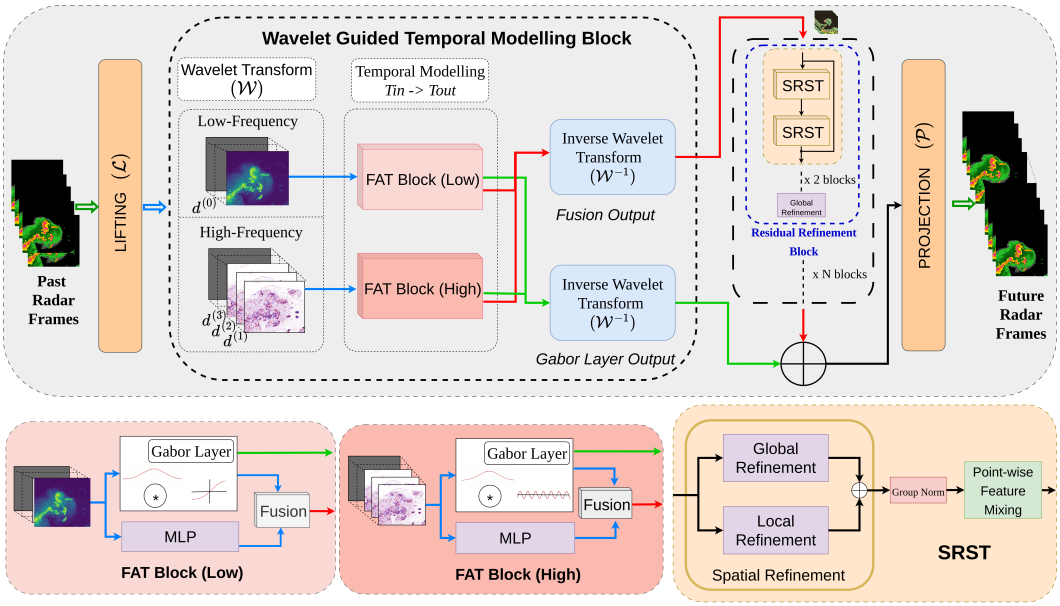


Figure 1: **DAWN-Cast.** Lifted frames are split by $\mathcal{W}$ into scale families, each forecast by its own FAT block (Gabor and MLP streams, fused). $\mathcal{W}^{-1}$ reconstructs the fused and Gabor outputs separately. The fused forecast is refined by RR blocks of SRST blocks, and the Gabor reconstruction is added as a residual before projection $\mathcal{P}$.

tion Kumar & Foufoula-Georgiou (1994); Yano & Jakubiak (2016); Kumar & Foufoula-Georgiou (1993). A J-level 2D wavelet transform of each radar frame yields one low-frequency band and three detail bands per level (Eq. 2); we call the low-frequency band, and the detail bands of each level, a *scale family*.

3.1.2 Why Gabor

Scale families do not evolve at the same pace: coarse-scale coefficients persist over many frames, whereas fine-scale coefficients decorrelate within a few (Appendix E). A temporal predictor should therefore be able to change its frequency and its selectivity from one scale to the next. An MLP does not offer this directly: its hidden units, shared by all predicted frames, apply a fixed pointwise activation with no explicit frequency or localization parameter. A Gabor activation does, since it multiplies a sinusoidal carrier by a Gaussian envelope, the classical localized, frequency-selective filter of signal analysis Daugman (1985); Dunn & Higgins (1995); Okajima (1998); Fathony et al. (2021).

For each scale family s (the low-frequency band and every high-frequency level), the Gabor stream maps the T_{in}-frame history $\mathbf{x} \in \mathbb{R}^{T_{in}}$ of a wavelet coefficient to T_{out} outputs, one Gabor neuron per predicted frame $\tau = 1, \ldots, T_{out}$:

$$G_{s,\tau}(\mathbf{x}) = \sin\Big(\lambda_s f_{s,\tau}\big(\mathbf{w}_{s,\tau}^{\top}\mathbf{x} + b_{s,\tau}\big)\Big)\, \exp\Big(-\tfrac{\gamma_{s,\tau}}{2}\,\|\mathbf{x} - \boldsymbol{\mu}_{s,\tau}\|_2^2\Big). \tag{1}$$

Here $\mathbf{w}_{s,\tau}$ and $b_{s,\tau}$ form a learned linear combination of the observed frames, $f_{s,\tau}$ is a learnable frequency scaled by a fixed family multiplier λ_s (a hyperparameter, Appendix K), $\boldsymbol{\mu}_{s,\tau} \in \mathbb{R}^{T_{in}}$ is a learned prototype history, and $\gamma_{s,\tau} > 0$ is a learnable bandwidth. Every neuron has its own parameters, so each future frame has its own frequency, input weighting, and history prototype. All parameters are shared across spatial positions and channels.

The two factors play distinct roles. The sinusoid sets how fast the response varies along $\mathbf{w}_{s,\tau}$, through the phase slope $\lambda_s f_{s,\tau}$. The Gaussian is a history-selective gate: with $\sigma_{s,\tau} = 1/\sqrt{\gamma_{s,\tau}}$, neuron τ responds strongly only when the observed history lies within about $\sigma_{s,\tau}$ of its prototype. As $\lambda_s f_{s,\tau} \to 0$ and $\gamma_{s,\tau} \to 0$, the neuron becomes a linear map of the history (up to a scale absorbed by the fusion layer, Eq. 3), so the Gabor stream contains the linear predictor as a special case.

3.2 Model Parts

3.2.1 Wavelet Guided Temporal Modelling (WGTM) Block

Scale families evolve at different paces (Sec. 3.1.2), so a single temporal mapping applied to a mixed-scale field must serve all of them at once. WGTM instead separates the scales first, in three steps (Fig. 1): it decomposes the input into scale families, forecasts every family with its own Frequency Adaptive Temporal (FAT) block, and reconstructs the predicted frames.

Decomposition (*wavelet transform*). WGTM receives the c-channel lifted features $Z \in \mathbb{R}^{T_{in}\times c\times H\times W}$. A J-level 2D discrete wavelet transform $\mathcal{W}_J$ is applied spatially to each frame separately. It returns one low-frequency band and three detail bands per level:

$$\mathcal{W}_J(Z) = \{S_J^{LL}\} \cup \bigcup_{j=1}^{J} \{S_j^{LH},\, S_j^{HL},\, S_j^{HH}\}. \tag{2}$$

We stack the three orientations of each level along the channel axis, giving $J{+}1$ *scale families*, the LL band and one detail family per level, indexed by $s \in \{\mathrm{LL}, 1, \ldots, J\}$. Family s has c_s channels ($c_{\mathrm{LL}} = c$, $c_s = 3c$ for detail families) and spatial size $H_s \times W_s$ ($H_J \times W_J$ for LL), with time moved to the last axis. Since the transform acts on space only, every coefficient remains a length-T_{in} time series on which the temporal blocks operate.

Frequency Adaptive Temporal (FAT) Block. Each family s has its own FAT block. The $J{+}1$ blocks share no parameters and exchange no information, and each applies the same weights at every spatial position and channel of its family. A FAT block maps the T_{in}-frame history x of a coefficient to T_{out} predicted frames through two streams: a *Gabor stream* $\mathrm{Gabor}_s(x)$, with one Gabor neuron per predicted frame (Eq. 1), and an *MLP stream* $\mathrm{MLP}_s(x)$, a two-layer perceptron with SELU activation along the time axis. Applying both streams to a family S_s gives $G_s = \mathrm{Gabor}_s(S_s)$ and $M_s = \mathrm{MLP}_s(S_s)$, both in $\mathbb{R}^{c_s\times H_s\times W_s\times T_{out}}$. A learned $1{\times}1{\times}1$ convolution ϕ_s then fuses them, projecting the $2c_s$ concatenated channels back to c_s:

$$F_s = \phi_s\big(\mathrm{Concat}[\,G_s,\, M_s\,]\big). \tag{3}$$

The fused output F_s carries the family's forecast, and the Gabor output G_s is also kept separately.

Reconstruction (*inverse wavelet transform*). The inverse transform $\mathcal{W}_J^{-1}$, applied to every predicted frame, recombines the fused outputs (Eq. 3) into $\hat{Z}_{\mathrm{fused}}$ and, separately, the Gabor outputs into a residual R_{Gabor}:

$$\hat{Z}_{\mathrm{fused}} = \mathcal{W}_J^{-1}\big(\{F_s\}\big), \qquad R_{\mathrm{Gabor}} = \mathcal{W}_J^{-1}\big(\{G_s\}\big), \tag{4}$$

both in $\mathbb{R}^{T_{out}\times c\times H\times W}$. The fused reconstruction enters the refinement stage, whereas R_{Gabor} bypasses it and is added to the refined output (Sec. 3.2.2).

3.2.2 Refinement Stage

WGTM leaves two gaps. First, the FAT blocks act at fixed positions with no spatial receptive field: they model each family's temporal statistics but cannot follow spatial displacement, which matters most at fine scales, whose coefficients decorrelate within 15 min at a fixed position (Appendix E). Second, each family is forecast independently, so the reconstructed field can be spatially inconsistent where the scales rejoin. We address both with a refinement stage of Residual Refinement (RR) blocks built from Spectral Refinement Spatio-Temporal (SRST) blocks (Fig. 1). It combines two known ideas, Fourier-domain mixing for global context Guibas et al. (2021) and a global spectral stream in parallel with a local convolutional one Wu et al. (2024). We claim not the streams themselves, but their role here: to repair what independent scale processing leaves behind.

SRST block (*joint space–time refinement*). We merge the temporal and channel axes of $\hat{Z}_{\mathrm{fused}}$, giving $U = \mathrm{Fold}(\hat{Z}_{\mathrm{fused}}) \in \mathbb{R}^{H\times W\times D}$ with $D = T_{out}\cdot c$, so that refinement sees space and time together.

Local-global spatial refinement runs two streams in parallel. The global stream applies AFNO-style Fourier-domain mixing Guibas et al. (2021) (a 2D real FFT over the spatial axes, a two-layer complex-valued transform, and an inverse FFT with a skip connection), so every output location can draw on the entire field. The local stream is a depthwise $k \times k$ convolution Chollet (2017), giving each temporal-channel map its own spatial kernel. Their sum is the block's spatial refinement:

$$\mathrm{Spec}(U) = \mathrm{AFNO}(U), \qquad \mathrm{Loc}(U) = \mathrm{DWConv}_{k\times k}(U), \qquad U_{\mathrm{spatial}} = \mathrm{Spec}(U) + \mathrm{Loc}(U). \tag{5}$$

The output U_{spatial} is normalized (GroupNorm, SiLU) and passed through point-wise feature mixing (PFM), a two-layer 1×1 convolution over the $D = T_{out} \cdot c$ channels that recombines information across predicted frames and channels at every pixel:

$$\mathrm{SRST}(U) = \mathrm{PFM}\Big(\mathrm{SiLU}\big(\mathrm{GN}(U_{\mathrm{spatial}})\big)\Big). \tag{6}$$

RR block. (*residual composition*). Two SRST blocks with a skip connection form one RR block,

$$\mathrm{RR}(U) = \mathrm{SRST}\big(\mathrm{SRST}(U)\big) + U. \tag{7}$$

The refinement stage stacks two RR blocks and closes with one further global spectral layer.

Gabor residual. The Gabor reconstruction R_{Gabor} from WGTM (Eq. 4) bypasses the refinement stage and is added to its output He et al. (2016), so the scale-specific temporal response of the Gabor streams reaches the output directly. The refined forecast is

$$\hat{Z} = \mathrm{Fold}^{-1}\Big(\mathrm{Spec}\big(\mathrm{RR}(\mathrm{RR}(U))\big)\Big) + R_{\mathrm{Gabor}} \in \mathbb{R}^{T_{out}\times c\times H\times W}, \tag{8}$$

which the projection then maps back to radar frames.

4 Experiments

4.1 Datasets

We evaluate on four radar precipitation datasets following the protocol of Yu et al. (2024); Lin et al. (2025): SEVIR (continental United States, NEXRAD VIL, 5-min) Veillette et al. (2020), MeteoNet (France, 5-min) Larvor et al. (2020), Shanghai Radar (Pudong, ∼6-min) Chen et al. (2020), and CIKM (Guangdong, 6-min) Yao & Li (2017). All frames are rescaled to 128×128, with input/output lengths of 5/20 for SEVIR, MeteoNet, and Shanghai, and 5/10 for CIKM. Splits and preprocessing are in Table 6 and Appendix B.

4.2 Implementation and Training

DAWN-Cast operates directly in pixel space. We train with AdamW, a learning rate of 1×10^{-4}, batch size 4, for 50 epochs, using the FACL loss Yan et al. (2024), on a single NVIDIA RTX A6000 GPU (48 GB VRAM). Wavelet configuration and Gabor initialization hyperparameters are selected per dataset by validation performance (Appendix K); a two-stage recipe that avoids this search at a small, dataset-dependent cost is evaluated in Appendix F. All reported models, including baselines, are single runs with seed 0, the seed on which hyperparameters were selected; five-seed statistics are in Appendix L and compute costs in Appendix J.

4.3 Quantitative Results

Table 1 compares DAWN-Cast against eleven baselines on SEVIR, MeteoNet, Shanghai, and CIKM; metric definitions are given in Appendix C.

DAWN-Cast achieves the best result on 19 of the 20 reported metrics and is second-best on the remaining one (CIKM CSI-40). It has the lowest MSE on all four datasets, which differ in region, sensor, and precipitation climatology, and its CSI-M margin over the strongest baseline is consistent (2.2–4.6%) even though that baseline changes from dataset to dataset (TrajGRU on SEVIR, AlphaPre on MeteoNet and Shanghai, EarthFormer on CIKM). Per-lead-time curves on CIKM (Appendix M) show that DAWN-Cast leads from the second to the ninth predicted frame, with the largest margin at intermediate lead times.

Table 1: Performance across four precipitation nowcasting benchmarks. We report CSI-M, two dataset-specific CSI thresholds, HSS, and MSE; best in **bold**, second-best underlined.

SEVIR

Model	CSI-M↑	CSI-181↑	CSI-219↑	HSS↑	MSE↓
ConvLSTM	0.3166	0.1499	0.0446	0.4036	462.42
TrajGRU	0.3676	0.2029	0.1154	0.4738	369.83
PhyDNet	0.3477	0.1863	0.0990	0.4505	400.84
MAU	0.3340	0.1603	0.0770	0.4312	401.84
SimVP	0.3579	0.1885	0.1057	0.4600	376.53
EarthFormer	0.3284	0.1465	0.0564	0.4244	373.64
EarthFarseer	0.3148	0.1381	0.0691	0.4080	429.62
WADEPre†	0.3524	0.1867	0.1032	0.4497	398.83
DiffCast	0.3153	0.1556	0.0843	0.4106	494.25
AlphaPre	0.3634	0.2033	0.0982	0.4639	373.25
exPreCast	0.3420	0.1694	0.0851	0.4386	387.06
DAWN-CAST	**0.3787**	**0.2177**	**0.1184**	**0.4847**	**340.90**

CIKM

Model	CSI-M↑	CSI-35↑	CSI-40↑	HSS↑	MSE↓
ConvLSTM	0.2758	0.1737	0.1066	0.3570	49.46
TrajGRU	0.3182	0.2116	0.1436	0.4107	37.01
PhyDNet	0.3108	0.2155	0.1354	0.4032	38.89
MAU	0.3105	0.2095	0.1362	0.4027	37.48
SimVP	0.3243	0.2305	0.1452	0.4190	40.31
EarthFormer	0.3260	0.2311	**0.1559**	0.4217	41.12
EarthFarseer	0.2970	0.1984	0.1272	0.3870	41.56
WADEPre†	0.2957	0.1908	0.1203	0.3825	36.19
DiffCast	0.3100	0.2238	0.1491	0.4069	45.31
AlphaPre	0.3096	0.2140	0.1408	0.3992	44.68
exPreCast	0.3169	0.2208	0.1421	0.4106	37.36
DAWN-CAST	**0.3411**	**0.2424**	0.1549	**0.4385**	**34.13**

MeteoNet

Model	CSI-M↑	CSI-24↑	CSI-32↑	HSS↑	MSE↓
ConvLSTM	0.4358	0.4282	0.2455	0.5716	10.12
TrajGRU	0.4155	0.4020	0.2340	0.5513	10.59
PhyDNet	0.4112	0.4033	0.2282	0.5504	10.74
MAU	0.4181	0.4029	0.2353	0.5575	9.89
SimVP	0.4404	0.4331	0.2621	0.5808	10.35
EarthFormer	0.3919	0.3847	0.2126	0.5342	12.73
EarthFarseer	0.3578	0.3320	0.1478	0.4909	11.52
WADEPre†	0.4200	0.4054	0.2388	0.5497	10.26
DiffCast	0.4100	0.3979	0.2520	0.5436	14.57
AlphaPre	0.4433	0.4320	0.2702	0.5791	9.84
exPreCast	0.3953	0.3794	0.2077	0.5289	11.37
DAWN-CAST	**0.4529**	**0.4449**	**0.2758**	**0.5890**	**9.31**

Shanghai

Model	CSI-M↑	CSI-35↑	CSI-40↑	HSS↑	MSE↓
ConvLSTM	0.4274	0.3945	0.2822	0.5652	29.58
TrajGRU	0.4182	0.3857	0.2712	0.5581	27.17
PhyDNet	0.4181	0.3864	0.2743	0.5599	29.32
MAU	0.4217	0.3860	0.2724	0.5638	29.03
SimVP	0.4341	0.3973	0.2799	0.5762	26.53
EarthFormer	0.4000	0.3679	0.2553	0.5427	32.68
EarthFarseer	0.3444	0.3007	0.1856	0.4797	43.30
WADEPre†	0.4189	0.3835	0.2648	0.5535	30.30
DiffCast	0.3854	0.3518	0.2554	0.5180	39.94
AlphaPre	0.4420	0.4069	0.2988	0.5803	26.23
exPreCast	0.4107	0.3737	0.2646	0.5505	29.56
DAWN-CAST	**0.4525**	**0.4186**	**0.3087**	**0.5918**	**25.33**

†Trained with its native multi-scale curriculum Liu et al. (2026), which supervises its approximation and detail outputs with separate losses and therefore cannot be replaced by FACL without altering the method. All other baselines are trained with FACL Yan et al. (2024), which improved over each baseline's native loss for the majority of models.

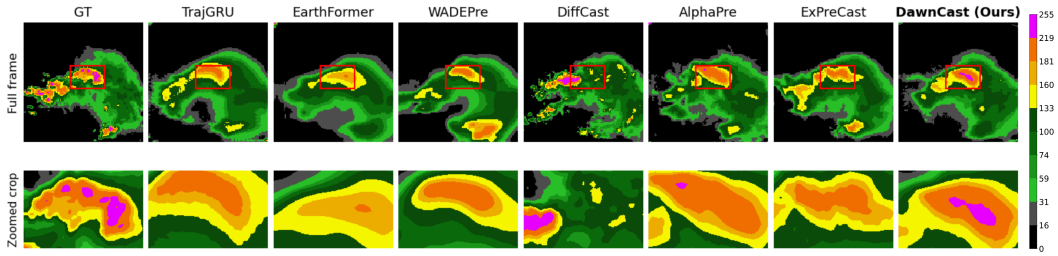


Figure 2: Qualitative comparison on SEVIR at the lead time of +100 min. Full frames (top) and zoomed crops (bottom) show that DAWN-CAST retains the high-intensity core (pink, $\geq$ 219) that competing models underestimate or smear out at this horizon.

4.4 QUALITATIVE RESULTS

At the final and most difficult lead time, DAWN-Cast keeps the SEVIR high-intensity core at the right location and magnitude (Fig. 2), whereas competing models underestimate or smear it, and on CIKM it preserves a coherent storm structure that baselines blur or fragment (Fig. 3). More examples are in Appendix N, and Appendix M relates these long-lead-time differences to per-frame scores.

5 MODEL ANALYSIS

5.1 WGTM BLOCK

In Table 2, *w/o WGTM* replaces wavelet decomposition and the per-family FAT blocks with a single MLP applied directly to the lifted features. WGTM improves 9 of the 10 reported metrics, with the largest relative gains at the highest intensity threshold (CSI-219 +16.8% on SEVIR, CSI-

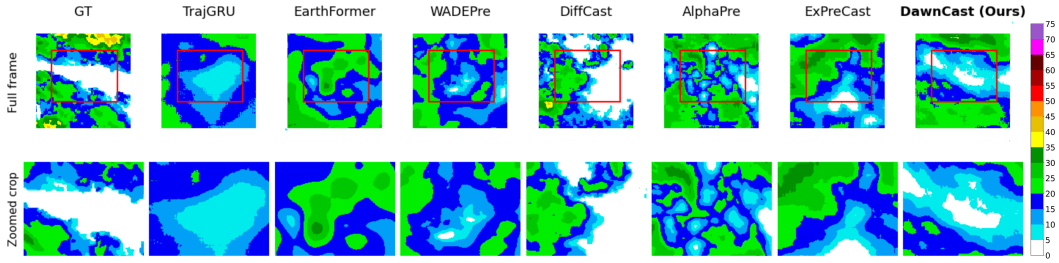


Figure 3: Qualitative comparison on CIKM at the lead time of $+60$ min. Full frames (top) and zoomed crops (bottom) show that DAWN-CAST preserves the overall storm structure that competing models blur or fragment at this horizon.

Table 2: **Temporal-path ablation on SEVIR and CIKM.** *w/o WGTM*: a single MLP on the lifted features, without wavelet decomposition; *Shared FAT block*: one FAT block for all scale families; *w/o FAT block*: per-family MLPs without the Gabor stream and its residual. Best in **bold**.

	SEVIR					CIKM				
Ablation	CSI-M↑	CSI-181↑	CSI-219↑	HSS↑	MSE↓	CSI-M↑	CSI-35↑	CSI-40↑	HSS↑	MSE↓
w/o WGTM block	0.3684	0.1936	0.1014	0.4674	**337.87**	0.3252	0.2205	0.1393	0.4154	35.47
Shared FAT block	0.3565	0.1815	0.0892	0.4530	354.29	0.3239	0.2190	0.1401	0.4151	35.81
w/o FAT block (Separate MLP)	0.3684	0.1989	0.1052	0.4690	348.36	0.3239	0.2193	0.1341	0.4143	34.14
Separate FAT blocks *(ours)*	**0.3787**	**0.2177**	**0.1184**	**0.4847**	340.90	**0.3411**	**0.2424**	**0.1549**	**0.4385**	**34.13**

Note. The shared FAT block processes all scale families jointly, which requires concatenating them along channels and hence equal spatial resolution. It therefore uses a level-1 decomposition ([LL, LH, HL, HH]). The decomposition is invertible, so no information is lost relative to $J{=}2$, but its LL band leaves coarse and intermediate scales unseparated.

40 $+11.2\%$ on CIKM). The exception, SEVIR MSE (340.90 vs. 337.87 without WGTM), reflects the usual trade-off between event-detection skill and pixel-averaged error: sharper, higher-intensity predictions can raise MSE slightly even as CSI and HSS improve (Subich et al., 2025).

The other two variants keep the decomposition and vary only the per-family temporal operator; separate FAT blocks beat both on all 10 metrics.

Figure 4 shows the same ordering qualitatively. On SEVIR, separation helps even without Gabor, and Gabor adds a further gain. On CIKM, separation alone brings little benefit (*w/o FAT* and *Shared FAT block* tie on CSI-M), and the gain comes from Gabor. *Shared FAT block* uses a level-1 decomposition, whereas our model uses $J{=}2$ on SEVIR and CIKM.

5.2 Residual Refinement Block

Table 3 isolates each refinement-stage component. Removing the entire stage causes the largest drop of any ablation in this paper (-21.9% CSI-M on MeteoNet, -16.8% on CIKM), so reconciling the independently evolved scales is essential. Local-global spatial refinement contributes most. Point-wise feature mixing matters more on MeteoNet, and the Gabor residual matters more on CIKM. Two effects are non-monotonic: removing local-global refinement alone gives a worse CIKM MSE (49.30) than removing the whole stage (43.05), and removing point-wise feature mixing slightly improves CIKM CSI-40 ($+1.7\%$).

5.3 Cross-Domain Generalization and Parameter-Efficient Adaptation

5.3.1 Zero-Shot Cross-Domain Generalization

For the zero-shot evaluation in Table 4, each dataset is normalized to $[0, 1]$ with its own intensity range (Appendix B); target frames are fed in normalized form, and predictions are mapped back to the target's range, where the target's own thresholds and MSE are computed with identical code for all models. Since SEVIR measures VIL and the targets measure reflectivity, this tests the transfer of

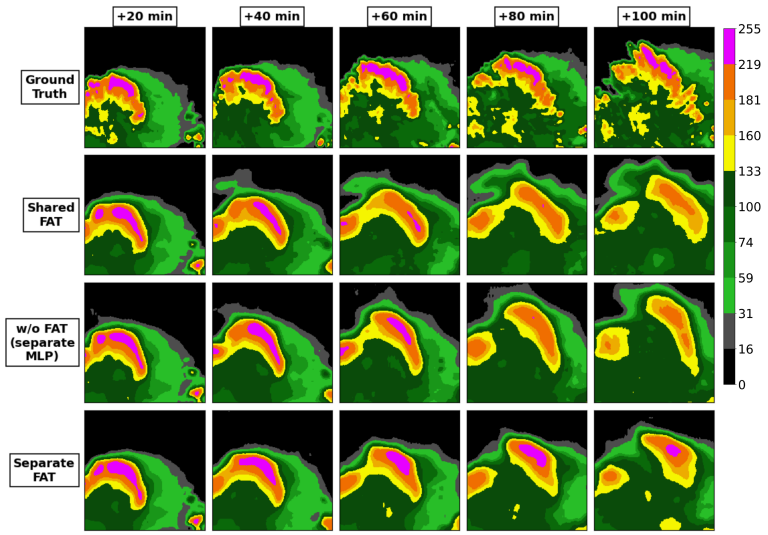


Figure 4: Qualitative ablation of the WGTM design on SEVIR. Rows differ in the temporal operator assigned per scale family. A single shared FAT block loses the high-intensity core by $+60$ min. Giving each family its own MLP (*w/o FAT*) extends this to $\sim+60$–80 min. Assigning each family its own full FAT block (*Separate FAT*, ours) retains the core through $+100$ min and most closely matches ground truth structure.

Table 3: **Refinement stage ablation on MeteoNet and CIKM.**

	MeteoNet					CIKM				
Ablation	CSI-M↑	CSI-24↑	CSI-32↑	HSS↑	MSE↓	CSI-M↑	CSI-35↑	CSI-40↑	HSS↑	MSE↓
w/o Residual Refinement block	0.3537	0.3363	0.1883	0.4825	14.28	0.2839	0.1723	0.1044	0.3640	43.05
w/o Local-global refinement block	0.4089	0.3901	0.2262	0.5384	10.47	0.3008	0.2048	0.1318	0.3853	49.30
w/o point-wise feature mixing	0.4331	0.4201	0.2586	0.5670	10.13	0.3356	0.2400	**0.1575**	0.4338	36.66
w/o Gabor residual	0.4401	0.4293	0.2464	0.5717	9.34	0.3279	0.2266	0.1456	0.4219	35.15
DAWN-Cast	**0.4529**	**0.4449**	**0.2758**	**0.5890**	**9.31**	**0.3411**	**0.2424**	0.1549	**0.4385**	**34.13**

normalized spatiotemporal dynamics rather than of physical calibration. DAWN-Cast is best on all 8 reported metrics, ahead of all eleven baselines. To isolate scale separation at matched parameter count, we compare against a shared-FAT variant of DAWN-Cast (59.51M vs. 59.54M parameters): it beats most baselines but loses three of four metrics to AlphaPre on both targets (winning only MSE), whereas the full model beats AlphaPre on every metric on both targets. Because the two variants have nearly identical parameter counts, the transfer gain comes from how the temporal path is organized rather than from capacity. WADEPre, the closest decomposition-based model, transfers worse than even the shared-FAT variant on 7 of 8 metrics. We therefore adapt this temporal path below. On Shanghai, the zero-shot model even has lower MSE than DAWN-Cast trained on Shanghai (23.40 vs. 25.33) while trailing it in CSI-M; Shanghai has the smallest training set (1,534 sequences vs. 23,808 for SEVIR), so we attribute this to SEVIR training acting as a strong prior. Zero-shot results on CIKM are in Appendix H.

5.3.2 Parameter-Efficient Adaptation

Table 5 fine-tunes only the lifting, projection, and temporal path (0.71% of parameters), keeping the refinement stage frozen at its SEVIR values. With 20% of target data, this adaptation outperforms full training on Shanghai and approaches it on MeteoNet (-3.15% CSI-M); with 100% data, it surpasses full training on every Shanghai metric and matches it on MeteoNet in CSI-M and HSS. Because it uses less target data, adaptation at 20% also reduces training time by 4.6–$5.3\times$ (Table 14, Appendix I). Unfreezing a wider surface instead (lifting, projection, and the second SRST block; 44.57% of parameters, Table 13) is slightly better on MeteoNet and slightly worse on Shanghai at

Table 4: **Zero-shot transfer from SEVIR to MeteoNet and Shanghai.** All models are trained on SEVIR and evaluated on the target test sets without target-domain data or parameter updates. Each model uses its best SEVIR checkpoint, trained as in Table 1.

		SEVIR → MeteoNet				**SEVIR → Shanghai**			
Model	Params	CSI-M↑	CSI-32↑	HSS↑	MSE↓	CSI-M↑	CSI-40↑	HSS↑	MSE↓
AlphaPre	89.03M	0.3650	0.2043	0.4937	12.23	0.4174	0.2853	0.5517	27.39
EarthFarseer	96.42M	0.3591	0.1879	0.4910	13.09	0.4020	0.2628	0.5391	29.56
SimVP	11.06M	0.3500	0.1916	0.4804	12.64	0.4069	0.2698	0.5421	27.66
TrajGRU	11.91M	0.3458	0.1668	0.4674	11.38	0.4059	0.2693	0.5393	27.44
WADEPre	30.42M	0.3368	0.1897	0.4536	12.49	0.4003	0.2675	0.5297	30.15
ConvLSTM	0.89M	0.3361	0.1815	0.4628	13.28	0.3904	0.2557	0.5223	31.63
ExPreCast	32.00M	0.3278	0.1732	0.4498	12.80	0.3839	0.2448	0.5143	28.48
DiffCast (PhyDNet)	49.36M	0.3158	0.1587	0.4393	16.33	0.3490	0.2170	0.4769	36.45
MAU	7.63M	0.3117	0.1456	0.4319	12.48	0.3874	0.2447	0.5207	26.77
EarthFormer	8.65M	0.3105	0.1501	0.4399	12.81	0.3616	0.2158	0.4967	28.07
PhyDNet	3.09M	0.3007	0.1433	0.4218	13.11	0.3641	0.2251	0.4955	30.05
DAWN-Cast-Shared FAT	59.51M	0.3565	0.1892	0.4792	10.61	0.4163	0.2734	0.5491	24.41
DAWN-Cast (ours)	59.54M	**0.3929**	**0.2159**	**0.5206**	**9.76**	**0.4422**	**0.3061**	**0.5786**	**23.40**

Table 5: **Parameter-efficient adaptation of the SEVIR DAWN-Cast to a new radar domain.** Only lifting, projection, and the temporal path are unfrozen (**423,009 of 59,541,089 parameters, 0.71%**); training settings as in Sec. 4.2. *Data* is the fraction of the target training set used (validation/test always full). Δ is relative to DAWN-Cast trained from scratch on that dataset; $+$ means adaptation outperforms it.

Dataset	Adaptation	Data	CSI-M↑	ΔCSI	HSS↑	ΔHSS	CSI-pool4↑	CSI-pool16↑	MSE↓
MeteoNet	Full-shot	100 %	0.4529	-	0.5890	-	0.5159	0.6132	9.31
	Zero-shot	0%	0.3929	-13.24%	0.5206	-11.61%	0.4310	0.4963	9.76
	Fine-tuned (0.71% params)	20%	0.4386	-3.15%	0.5759	-2.22%	0.4949	0.5671	9.95
	Fine-tuned (0.71% params)	50%	0.4476	-1.17%	0.5859	-0.52%	0.5100	0.5823	9.71
	Fine-tuned (0.71% params)	100%	0.4527	-0.05%	0.5910	+0.34%	0.5146	0.5896	9.68
Shanghai	Full-shot	100 %	0.4525	-	0.5918	-	0.5113	0.5842	25.33
	Zero-shot	0%	0.4422	-2.27%	0.5786	-2.23%	0.4925	0.5727	23.40
	Fine-tuned (0.71% params)	20%	0.4639	+2.51%	0.6037	+2.02%	0.5224	0.5946	23.28
	Fine-tuned (0.71% params)	50%	0.4699	+3.84%	0.6096	+3.00%	0.5249	0.5809	23.22
	Fine-tuned (0.71% params)	100%	0.4703	+3.94%	0.6112	+3.27%	0.5295	0.5893	23.08

20% and 50% data: adapting about 63× more parameters in the refinement stage does not improve on adapting the scale-specific temporal path alone.

6 Conclusion

We introduced DAWN-Cast, which assigns each wavelet scale family its own Frequency Adaptive Temporal block instead of sharing one temporal operator across scales, and reconciles the independently evolved scales with a spectral–local refinement stage. DAWN-Cast achieves the best result on 19 of 20 metrics across four radar datasets. Ablations show that per-family FAT blocks outperform shared and MLP-only variants and that the refinement stage is essential. The same organization transfers zero-shot to unseen radar domains better than eleven baselines and a parameter-matched shared-FAT variant, and adapts to a new domain by updating only 0.71% of its parameters.

AI USE STATEMENT

In this work, we used generative AI tools solely for language editing and polishing the manuscript, including improving grammar, clarity, readability, and sentence structure. Generative AI tools were not used to generate research ideas, develop the methodology, conduct experiments, analyze results, produce scientific claims, or create figures. All AI-assisted edits were carefully reviewed by the authors to ensure that they preserved the intended technical meaning and accurately reflected the work performed. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the source code used to implement and evaluate `DAWN-Cast` as supplementary material. The main paper describes the proposed architecture, training procedure, evaluation protocol, and baseline comparisons, while the appendix provides additional details on dataset preprocessing, evaluation metrics, model components, hyperparameter configurations, computational cost, statistical analysis, and cross-domain adaptation settings. Together with the supplementary code, these materials are intended to enable reproduction of the reported experiments and results.

REFERENCES

Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xiang Xinguang, and Wen Gao. Mau: A motion-aware unit for video prediction and beyond. *Advances in Neural Information Processing Systems*, 34:26950–26962, 2021.

Lei Chen, Yuan Cao, Leiming Ma, and Junping Zhang. A deep learning-based methodology for precipitation nowcasting with radar. *Earth and Space Science*, 7(2):e2019EA000812, 2020.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1800–1807. Ieee, 2017.

John G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, 1985. doi: 10.1364/JOSAA.2.001160.

D. Dunn and W. E. Higgins. Optimal gabor filters for texture segmentation. *IEEE Transactions on Image Processing*, 4(7):947–964, 1995. doi: 10.1109/83.392336.

Rizal Fathony, Anit Kumar Sahu, Devin Willmott, and J Zico Kolter. Multiplicative filter networks. *ICLR*, 2(3):5, 2021.

Christopher A Field and Alan H Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(3):369–390, 2007.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3170–3180, 2022a.

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

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:78621–78656, 2023.

Urs Germann and Isztar Zawadzki. Scale-dependence of the predictability of precipitation from continental radar images. part i: Description of the methodology. *Monthly Weather Review*, 130 (12):2859–2873, 2002.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11474–11484, 2020.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Praveen Kumar and Efi Foufoula-Georgiou. A multicomponent decomposition of spatial rainfall fields: 1. segregation of large-and small-scale features using wavelet transforms. *Water Resources Research*, 29(8):2515–2532, 1993.

Praveen Kumar and Efi Foufoula-Georgiou. Wavelet analysis in geophysics: An introduction. *Wavelets in geophysics*, 4:1–43, 1994.

Gwennaëlle Larvor, Léa Berthomier, Vincent Chabot, Brice Le Pape, Bruno Pradel, and Lior Perez. Meteonet, an open reference weather dataset by meteo-france. 2020. *URL https://www. kaggle. com/datasets/katerpillar/meteonet*, 2020.

Kenghong Lin, Baoquan Zhang, Demin Yu, Wenzhi Feng, Shidong Chen, Feifan Gao, Xutao Li, and Yunming Ye. Alphapre: Amplitude-phase disentanglement model for precipitation nowcasting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17841–17850, 2025.

Baitian Liu, Haiping Zhang, Huiling Yuan, Dongjing Wang, Ying Li, Feng Chen, and Hao Wu. Wadepre: A wavelet-based decomposition model for extreme precipitation nowcasting with multi-scale learning. *arXiv preprint arXiv:2602.02096*, 2026.

Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pp. 674–679, 1981.

Guy P Nason and Bernard W Silverman. The stationary wavelet transform and some statistical applications. In *Wavelets and statistics*, pp. 281–299. Springer, 1995.

K. Okajima. The gabor function extracts the maximum information from input local signals. *Neural Networks*, 11(3):435–439, 1998. doi: 10.1016/S0893-6080(98)00008-2.

Seppo Pulkkinen, Daniele Nerini, Andrés A Pérez Hortal, Carlos Velasco-Forero, Alan Seed, Urs Germann, and Loris Foresti. Pysteps: An open-source python library for probabilistic precipitation nowcasting (v1. 0). *Geoscientific Model Development*, 12(10):4185–4219, 2019.

Basivi Radhakrishna, Isztar Zawadzki, and Frédéric Fabry. Predictability of precipitation from continental radar images. part v: Growth and decay. *Journal of the Atmospheric Sciences*, 69(11): 3336–3349, 2012.

AW Seed. A dynamic and spatial scaling approach to advection forecasting. *Journal of Applied Meteorology*, 42(3):381–388, 2003.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in neural information processing systems*, 30, 2017.

Changhoon Song, Teng-Yuan Chang, and Youngjoon Hong. Extreme weather nowcasting via local precipitation pattern prediction. In *International Conference on Learning Representations*, volume 2026, pp. 109146–109177, 2026.

Christopher Subich, Syed Zahid Husain, Leo Separovic, and Jing Yang. Fixing the double penalty in data-driven weather forecasting through a modified spherical harmonic loss function. *arXiv preprint arXiv:2501.19374*, 2025.

Juanzhen Sun, Ming Xue, James W Wilson, Isztar Zawadzki, Sue P Ballard, Jeanette Onvlee-Hooimeyer, Paul Joe, Dale M Barker, Ping-Wah Li, Brian Golding, et al. Use of nwp for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society*, 95(3):409–426, 2014.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. 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.

Hao Wu, Yuxuan Liang, Wei Xiong, Zhengyang Zhou, Wei Huang, Shilong Wang, and Kun Wang. Earthfarsser: Versatile spatio-temporal dynamical systems modeling in one model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 15906–15914, 2024.

Chiu-Wai Yan, Shi Q Foo, Van H Trinh, Dit-Yan Yeung, Ka-Hing Wong, and Wai-Kin Wong. Fourier amplitude and correlation loss: Beyond using l2 loss for skillful precipitation nowcasting. *Advances in Neural Information Processing Systems*, 37:100007–100041, 2024.

Jun-Ichi Yano and Bogumil Jakubiak. Wavelet-based verification of the quantitative precipitation forecast. *Dynamics of Atmospheres and Oceans*, 74:14–29, 2016.

Yichen Yao and Zhongjie Li. Cikm analyticup 2017: short-term precipitation forecasting based on radar reflectivity images. In *Proceedings of the Conference on Information and Knowledge Management, Short-Term Quantitative Precipitation Forecasting Challenge, Singapore*, pp. 6–10, 2017.

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

A Appendix

B Dataset and Preprocessing Details

SEVIR dataset Veillette et al. (2020) is a curated, labeled dataset that synchronizes several data modalities, including visible satellite imagery, infrared satellite imagery (mid-level water vapor and clean longwave window channels), NEXRAD radar mosaics of vertically integrated liquid (VIL), and ground-based lightning detections. It contains 20,393 distinct weather events recorded between 2017 and 2020, with each event consisting of a 4-hour sequence of images sampled every 5 minutes over a 384 km × 384 km region within the continental U.S. For the task of forecasting future VIL frames up to 20 time steps (100 minutes) from 5 input frames (25 minutes), we extract sequences of 25 consecutive frames using a stride of 13 for each event. The dataset is partitioned into training, validation, and test subsets based on temporal cutoffs of January 1, 2019, and June 1, 2019. All frames are linearly rescaled to the range 0–255 and then binarized using the thresholds [16, 74, 133, 160, 181, 219] to compute CSI and HSS.

METEONET dataset Larvor et al. (2020) is a multimodal dataset that includes complete time series of satellite and radar imagery, numerical weather prediction outputs, and ground-based measurements. It covers a 550 km × 550 km region in northwestern France and spans a three-year period (2016–2018), with data recorded every 5 minutes.Adopting the same protocol as SEVIR, we follow Yu et al. (2024) and Lin et al. (2025) to partition the radar sequences collected from 2016 to 2018. For the CSI and HSS metrics, we adopt thresholds of [12, 18, 24, 32] following Yu et al. (2024) Lin et al. (2025).Detailed dataset statistics are summarized in Table 6.

Shanghai Radar dataset Chen et al. (2020) comprises continuous radar echo frames produced by volumetric scans conducted roughly every 6 minutes between October 2015 and July 2018 in Pudong, Shanghai. Each radar reflectivity map spans an area of 501 km × 501 km. We preprocess the echo sequences following the procedures in Yu et al. (2024) Lin et al. (2025). The pixel value range of the frames is constrained to [0–70], and thresholds of [20, 30, 35, 40] are used for calculating CSI and HSS.

CIKM dataset Yao & Li (2017) originates from the CIKM AnalytiCup 2017 Competition and captures precipitation events over a 101 km × 101 km region in Guangdong, China. Each data sample consists of 15 past radar echo frames, with a 6-minute interval between successive frames. We followed the preprocessing protocol of Lin et al. (2025) Yu et al. (2024), we pad each radar echo frame to 128 × 128 pixels and partition the dataset into training, validation, and test subsets in accordance with the original configuration. Pixel intensities in each frame are converted to reflectivity values in the range [0, 76] dBZ, and thresholds of [20, 30, 35, 40] are employed to calculate CSI and HSS. Detailed dataset statistics are summarized in Table 6.

Table 6: Data split statistics. N_{train}, N_{valid} and N_{test} refer to the number of samples allocated to the training, validation, and test splits.

Dataset	N_{train}	N_{valid}	N_{test}	(C, H, W)	T_i	T_o
Shanghai	1534	526	526	(1,128,128)	5	20
MeteoNet	6308	1310	1310	(1,128,128)	5	20
CIKM	8000	2000	4000	(1,128,128)	5	10
SEVIR	23808	6016	8100	(1,128,128)	5	20

C Evaluation Metrics

All models are evaluated on the test split at 128×128 with the identical evaluation code, from the *best* checkpoint of each run – the checkpoint with the highest validation CSI-M, selected by validating every 5 epochs. Let H, M and F denote hits, misses and false alarms obtained by binarising prediction and ground truth at an intensity threshold.

- **CSI** $= H/(H + M + F)$, the Critical Success Index, computed per lead time and averaged over all lead times.
- **CSI-M** $\uparrow$ – CSI averaged over *all* intensity thresholds of the dataset (SEVIR: 16, 74, 133, 160, 181, 219; MeteoNet: 12, 18, 24, 32; Shanghai and CIKM: 20, 30, 35, 40). This is the aggregate skill score and the quantity used for checkpoint selection.
- **CSI-H2 / CSI-H** $\uparrow$ – CSI at the second-highest and highest thresholds of each dataset, i.e. the extreme-intensity regime; the main tables name them by threshold (SEVIR: 181/219, MeteoNet: 24/32, Shanghai and CIKM: 35/40).
- **CSI-pool4 / CSI-pool16** $\uparrow$ – CSI computed after 4×4 and 16×16 max-pooling of both fields. Pooling relaxes exact pixel alignment, so these measure whether precipitation is placed in approximately the right region rather than the exact right pixel.
- **HSS** $\uparrow$ – Heidke Skill Score, skill relative to a random forecast, corrected for chance agreement. Like CSI-M, it is averaged over all thresholds of the dataset (HSS-M) and over lead times, unless reported per lead time.
- **SSIM** $\uparrow$ – Structural Similarity Index; structural and textural fidelity of the predicted fields.
- **PSNR** $\uparrow$ – Peak Signal-to-Noise Ratio in dB.
- **MSE** $\downarrow$ – mean squared error in scaled pixel space.

D Wavelet Interpretability

Figure 5 presents a dataset-wide statistical validation of the wavelet scale separation underlying DAWN-Cast. For each of the four datasets: SEVIR, MeteoNet, CIKM, and Shanghai, we compute the lag-1 spatial autocorrelation ρ and the mean normalized radial Power Spectral Density (PSD) across all precipitation events, reporting ensemble statistics over the full test split.

Spatial autocorrelation (top row). The boxplots reveal a consistent and dataset-independent pattern. The ground truth (GT) and the LL subband both exhibit high spatial autocorrelation ($\rho > 0.95$ in all four datasets), confirming that the low-frequency approximation faithfully preserves the coherent large-scale structure of the precipitation field. In contrast, the high-frequency detail subbands

(LH, HL, HH) universally exhibit near-zero or negative autocorrelation ($\rho \leq 0$), characteristic of spatially decorrelated signals. This is partly expected by construction, since detail coefficients are outputs of high-pass filters; the analysis therefore confirms that the decomposition separates spatial scales, not that the scales evolve differently in time, which Appendix E tests directly. The interquartile ranges of the LL subband remain narrow and well above zero in all cases, while those of the high-frequency subbands remain tightly clustered around zero, indicating that this scale separation is not only consistent on average but also stable across individual precipitation events.

Normalized Power Spectral Density (bottom row). The radial PSDs provide a complementary spectral view of the same separation. The LL subband (blue) concentrates its energy at low wavenumbers and closely follows the ground truth spectrum (black) across the entire low-wavenumber regime, confirming that the approximation subband captures the dominant large-scale spectral content of the precipitation field. The high-frequency subbands (red) exhibit a markedly different spectral profile: their energy is concentrated at high wavenumbers and decays steeply, with a significantly shallower slope than the LL subband corresponding to fine-scale variability. The shaded regions (± 1 standard deviation across events) are narrow for the LL subband and wider for the high-frequency subbands, reflecting the greater event-to-event variability in variable fine-scale structure compared to the more predictable large-scale convective organization. Together, the autocorrelation and PSD analyses confirm that 2D discrete wavelet decomposition induces a meaningful scale separation in precipitation fields, on which the per-family temporal modeling of the WGTM block operates.

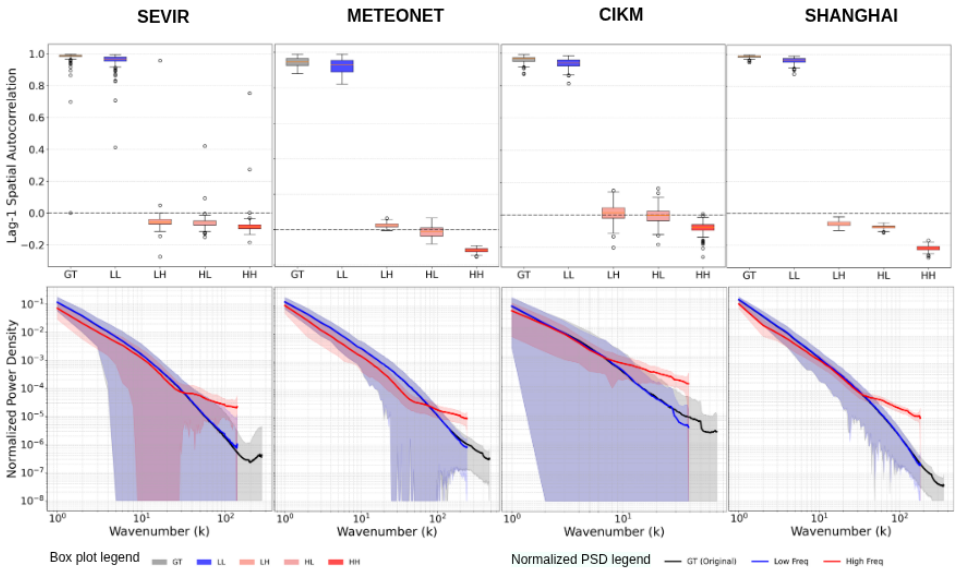


Figure 5: Statistical validation of wavelet-based convective scale separation across diverse meteorological datasets. Ensemble analysis of spatial autocorrelation (*top row*) and mean normalized Power Spectral Density (*bottom row*) across precipitation events from SEVIR, CIKM, Shanghai, and MeteoNet. Boxplots demonstrate that the LL subband consistently maintains structural continuity ($\rho > 0.95$) across all climatologies, whereas high-frequency subbands exhibit near-zero or negative spatial autocorrelation ($\rho \leq 0$). Here ρ indicates lag-1 spatial autocorrelation. The ensemble PSDs (shaded regions represent ± 1 standard deviation) confirm a robust, dataset-independent decoupling of macro-scale convective flows from localized high-frequency perturbations.

E SCALE-DEPENDENT EVOLUTION OF WAVELET SUBBANDS

A premise of DAWN-Cast is that low- and high-frequency wavelet subbands of precipitation fields evolve differently in time. This appendix tests that premise directly on SEVIR (Veillette et al., 2020). We follow the Lagrangian-persistence framework of Germann & Zawadzki (2002), which

separates the motion of precipitation patterns from their growth and decay, and apply it to each wavelet subband.

Data. We use SEVIR vertically integrated liquid (VIL) sequences at 384×384 pixels (1 km) and 5-min resolution, scaled to $[0, 1]$. Each forecast origin t_0 is kept only if at least 5% of its pixels exceed a VIL level of 16 (on the 0–255 scale). This yields 23,857 origins from 3,165 events: 593 storm events, which are matched to NOAA Storm Events reports, and 2,572 random events. For stratification, we group storm event types into *convective* (Hail, Tornado, Thunderstorm Wind, Funnel Cloud, Lightning; 500 events) and *hydro* (Flash Flood, Flood, Heavy Rain; 93 events). This grouping is our own.

Subband decomposition. Each frame is decomposed with a four-level orthonormal stationary (undecimated) wavelet transform (Nason & Silverman, 1995) using the `db2` wavelet. This gives an approximation band A_4 (scales > 32 km) and detail bands D_4 (16–32 km), D_3 (8–16 km), D_2 (4–8 km) and D_1 (2–4 km). Horizontal, vertical and diagonal coefficients are pooled within each level. We use the undecimated transform because a decimated transform is not shift-invariant: pure translation would then alter fine-scale coefficients and inflate their apparent decorrelation.

Eulerian and Lagrangian persistence. A motion field $\mathbf{V}$ is estimated from frames $t_0 - 2, \ldots, t_0$ with the dense Lucas–Kanade method (Lucas & Kanade, 1981) of `pysteps` (Pulkkinen et al., 2019). $\mathbf{V}$ is held fixed over the forecast. For lead times $\tau_k = 5k$ min, $k = 1, \ldots, 12$, each subband at t_0 is advected along backward semi-Lagrangian trajectories and interpolated once per lead with cubic splines. This gives $\mathcal{A}_k[W_j(t_0)]$. For band j we compute:

$$\rho_j^E(k) = \operatorname{corr}\big(W_j(t_0),\, W_j(t_0 + k)\big), \tag{9}$$

$$\rho_j^L(k) = \operatorname{corr}\big(\mathcal{A}_k[W_j(t_0)],\, W_j(t_0 + k)\big), \tag{10}$$

$$R_j(k) = \frac{\big\|W_j(t_0 + k) - \mathcal{A}_k[W_j(t_0)]\big\|^2}{\big\|W_j(t_0 + k)\big\|^2}. \tag{11}$$

The Lagrangian correlation ρ^L removes the effect of motion, so its decay with lead time measures non-advective change. R_j is the fraction of band energy not explained by advection, in the spirit of the growth-and-decay fields of Radhakrishna et al. (2012). If the prediction and the truth are uncorrelated and have equal energy, $R_j = 2$. A 16-pixel border and trajectories that leave the domain are excluded.

Pure-advection control. Motion-estimation and interpolation errors also reduce ρ^L, especially at fine scales. To quantify this, we build a surrogate sequence by advecting frame $t_0 - 2$ with the estimated $\mathbf{V}$. This surrogate contains no growth or decay. We then run it through the identical pipeline, including re-estimation of the motion field. Any loss of ρ^L in the surrogate is therefore attributable to the method rather than to atmospheric evolution.

Lagrangian lifetimes. We define the control-corrected ratio $r_j(k) = \bar{\rho}^L_{j,\mathrm{real}}(k)/\bar{\rho}^L_{j,\mathrm{ctrl}}(k)$ and assume $r_j(k) = \exp(-\tau_k/T_j)$. The lifetime T_j is fitted by least squares through the origin in $\ln r_j$. The fit uses leads with $0.1 < r_j < 0.98$: ratios near one carry no decay information, and ratios near zero are dominated by noise. A scaling exponent z in $T \propto L^z$ is fitted over D_4–D_2, with L the geometric centre of each band's scale range. All 95% confidence intervals come from a cluster bootstrap over events (Field & Welsh, 2007) with 1,000 resamples, so that overlapping origins from the same event are resampled together.

Results. Figure 6 and Table 7 show the main result. After motion is removed, subbands lose correlation in strict order of scale at every lead time and in every subset. At 15 min, ρ^L falls from 0.95 (A_4) to 0.66, 0.43, 0.21 and 0.06 (D_4 to D_1). The surrogate stays close to one for A_4 to D_3. The gap between real data and surrogate, which is the loss attributable to non-advective evolution, therefore grows toward fine scales: 0.05, 0.31, 0.51 and 0.66 for A_4 to D_2. Without motion correction, all detail bands are already decorrelated at 15 min ($\rho^E \leq 0.20$), in both real data and surrogate. At a fixed position, fine-scale structure is therefore largely unpredictable from its own history; forecasting it requires spatial context, which DAWN-Cast supplies in its refinement stage (Sec. 3.2.2). The

Table 7: Eulerian correlation ρ^E, Lagrangian correlation ρ^L and residual energy fraction R per subband, over all events (23,857 origins). "Control" is the pure-advection surrogate processed identically.

		15 min			30 min			60 min		
	Band	ρ^E	ρ^L	R	ρ^E	ρ^L	R	ρ^E	ρ^L	R
Real	A_4 (>32 km)	0.85	0.95	0.05	0.68	0.87	0.13	0.47	0.70	0.30
	D_4 (16–32 km)	0.20	0.66	0.64	0.05	0.36	1.28	0.02	0.10	1.87
	D_3 (8–16 km)	0.07	0.43	1.09	0.03	0.16	1.69	0.01	0.02	2.00
	D_2 (4–8 km)	0.04	0.21	1.54	0.02	0.06	1.90	0.02	0.01	2.03
	D_1 (2–4 km)	0.03	0.06	1.73	0.02	0.01	1.83	0.01	0.00	1.87
Control	A_4	0.88	1.00	0.01	0.73	0.99	0.01	0.53	0.97	0.03
	D_4	0.24	0.97	0.04	0.07	0.93	0.11	0.02	0.83	0.29
	D_3	0.09	0.94	0.08	0.02	0.86	0.22	0.00	0.68	0.55
	D_2	0.03	0.87	0.20	0.01	0.70	0.49	0.00	0.45	0.98
	D_1	0.01	0.67	0.55	0.00	0.42	1.00	0.00	0.20	1.44

residual energy R_j already exceeds one at 15 min for D_3–D_1 and approaches two at longer leads. The approximation band remains largely advective ($R = 0.30$ at 60 min).

Table 8 and Figure 7 report lifetimes. Scales above 32 km persist far beyond the 60-min analysis window (the fitted ≈3 h is an extrapolation). Lifetime decreases from 29.3 min at 16–32 km to 11.0 min at 4–8 km, following $T \propto L^{0.70}$ (95% CI $[0.70, 0.71]$). Halving the spatial scale thus shortens the lifetime by a factor of about $2^{0.7} \approx 1.6$. This is consistent with the scale-dependent lifetimes underlying spectral nowcasting schemes (Seed, 2003). The exponent is stable across event types (0.68–0.72) and is similar without the control correction (0.75). Convective events, however, have uniformly shorter lifetimes than random events: by about 25% for A_4 and about 14% for D_3–D_2, with non-overlapping confidence intervals. Convection thus shifts lifetimes downward at all scales without changing how lifetime scales with size.

Limitations. (i) The lifetime of D_1 (2–4 km) is not resolved. Its fitted value (≈ 4.6 min) rests on two leads, falls below the 5-min frame interval, and the surrogate itself loses substantial correlation in this band. We therefore make no claim for D_1. This is consistent with Germann & Zawadzki (2002), who likewise draw no conclusions for scales below about 10 km. (ii) Decay is faster than exponential: lifetimes estimated at single leads decrease with lead time. T_j should therefore be read as an effective e-folding time over 5–60 min. (iii) z is fitted over three octaves. Its confidence interval reflects sampling uncertainty only, not uncertainty in the power-law form. (iv) Motion is assumed stationary over the 60-min horizon.

Relation to DAWN-Cast's scale families. This analysis uses raw VIL at 1 km, whereas DAWN-Cast applies a decimated transform to lifted features at 128×128, i.e. about 3 km per pixel on SEVIR. At that resolution, the model's level-1 and level-2 detail families span roughly 6–12 km and 12–24 km, the D_3–D_4 range above, with lifetimes of 18–29 min (about 4–6 frames), whereas its approximation band (>24 km) persists far longer. The contrast in persistence therefore holds at the scales the model actually separates.

Implication. Over a 1-h nowcast, the approximation band is largely predictable by advection, D_4 retains skill for roughly half an hour, and D_3 and finer lose most of their advective predictability within 15–30 min. This observation motivates our central design choice: a single shared temporal operator would have to fit persistence times that differ by an order of magnitude across scales, so DAWN-Cast assigns each scale family an independent temporal operator. Because fine-scale coefficients also decorrelate quickly at fixed positions, the temporal operators are complemented by a spatial refinement stage.

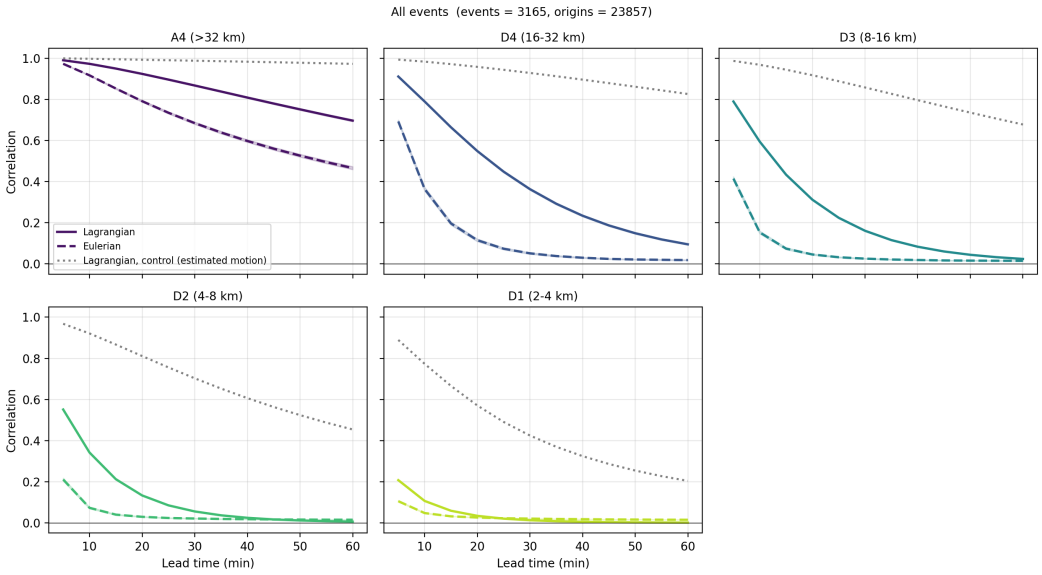


Figure 6: Eulerian (dashed) and Lagrangian (solid) correlation per wavelet subband against lead time, over all 3,165 SEVIR events. Shading shows 95% cluster-bootstrap intervals over events. The dotted grey line is the Lagrangian correlation of the pure-advection control, which has no growth or decay. Its departure from one measures the pipeline's own error.

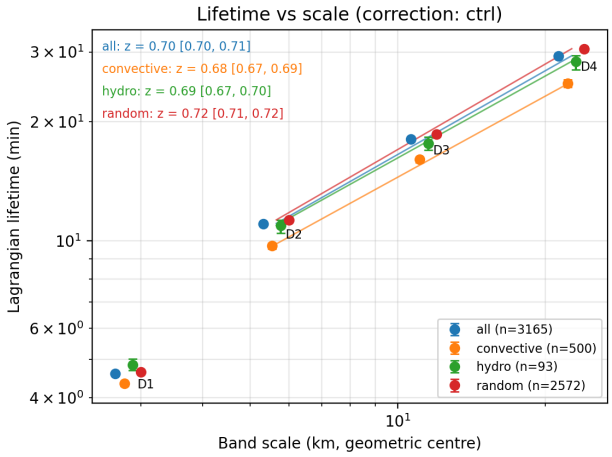


Figure 7: Control-corrected Lagrangian lifetime against subband scale (log–log), by event category, with 95% cluster-bootstrap intervals. Lines show power-law fits $T \propto L^z$ over D_4–D_2.

F TWO-STAGE TRAINING.

The main results select the wavelet configuration and the Gabor initialization (family multipliers λ_s, initial bandwidths, and initial weight scales) per dataset by validation CSI-M (Appendix K). To test whether DAWN-Cast's gains depend on this search, we evaluate a two-stage recipe that replaces the search over Gabor hyperparameters.

Procedure. *Stage 1* trains only the lifting, the wavelet temporal path (Gabor and MLP streams with their fusion), and the projection, without the refinement stage, from a single default Gabor initialization shared by all datasets (λ_s=1 for every scale family, $f_{s,\tau} \sim U(0,1)$, $\gamma_{s,\tau} \sim \mathrm{Gamma}(1,1)$, weight scale 0.1). Because the temporal path is then the only route to the output, the Gabor streams

Table 8: Control-corrected Lagrangian lifetimes T_j (min) with 95% cluster-bootstrap intervals, and scaling exponent z ($T \propto L^z$, fitted over D_4–D_2). D_1 is not resolved (see text). Without the control correction, $z = 0.75\,[0.75, 0.76]$ for all events.

Band	All (3,165)	Convective (500)	Hydro (93)	Random (2,572)
A_4 (>32 km)	197.4 [193.9, 201.0]	156.3 [151.0, 162.3]	197.3 [179.6, 215.9]	207.4 [203.3, 211.9]
D_4 (16–32 km)	29.3 [28.9, 29.6]	25.0 [24.5, 25.5]	28.4 [27.0, 29.4]	30.6 [30.2, 30.9]
D_3 (8–16 km)	18.0 [17.9, 18.2]	16.0 [15.7, 16.2]	17.6 [16.9, 18.3]	18.6 [18.4, 18.8]
D_2 (4–8 km)	11.0 [11.0, 11.1]	9.7 [9.5, 9.9]	10.9 [10.4, 11.3]	11.3 [11.2, 11.4]
D_1 (2–4 km)	n.r.	n.r.	n.r.	n.r.
z	0.70 [0.70, 0.71]	0.68 [0.67, 0.69]	0.69 [0.67, 0.70]	0.72 [0.71, 0.72]

Table 9: Two-stage training (Stage 1: lifting + wavelet temporal path + projection; Stage 2: full DAWN-Cast with the Gabor initialised from Stage 1) compared against the published DAWN-Cast best model on the four pixel-space datasets. All runs are 128×128, T_{in}=5, 50 epochs, FACL objective only, evaluated on the test split with the best-validation checkpoint. Δ is the percentage change of two-stage over the published best; for MSE a positive Δ denotes a reduction in error. CSI-H and CSI-H2 are the two highest intensity thresholds of each dataset.

Dataset	Model	CSI-M↑	CSI-H2↑	CSI-H↑	CSI-pool4↑	CSI-pool16↑	HSS↑	SSIM↑	MSE↓
SEVIR	DAWN-Cast (Manual Tuning)	0.3787	0.2177	0.1184	0.4211	0.4862	0.4847	0.6821	340.90
	Two-stage (ours)	0.3729	0.2102	0.1139	0.4121	0.4665	0.4770	0.6727	337.06
	Δ vs best	-1.54%	-3.43%	-3.81%	-2.15%	-4.04%	-1.58%	-1.37%	+1.13%
MeteoNet	DAWN-Cast (Manual Tuning)	0.4529	0.4449	0.2758	0.5159	0.6132	0.5890	0.8419	9.31
	Two-stage (ours)	0.4494	0.4393	0.2774	0.5211	0.6299	0.5864	0.8427	9.51
	Δ vs best	-0.78%	-1.25%	+0.58%	+1.02%	+2.73%	-0.45%	+0.10%	-2.13%
Shanghai	DAWN-Cast (Manual Tuning)	0.4525	0.4186	0.3087	0.5113	0.5842	0.5918	0.7301	25.33
	Two-stage (ours)	0.4518	0.4186	0.3113	0.5152	0.5835	0.5916	0.7018	25.75
	Δ vs best	-0.15%	-0.00%	+0.84%	+0.77%	-0.13%	-0.03%	-3.88%	-1.66%
CIKM	DAWN-Cast (Manual Tuning)	0.3411	0.2424	0.1549	0.3737	0.4282	0.4385	0.6087	34.13
	Two-stage (ours) (best of seeds 0–4)	0.3313	0.2338	0.1646	0.3653	0.4234	0.4294	0.6589	34.67
	Δ vs best	-2.88%	-3.56%	+6.26%	-2.24%	-1.12%	-2.07%	+8.24%	-1.59%

must carry the forecast on their own, and their frequencies and bandwidths adapt to each dataset's scale families. *Stage 2* trains the full DAWN-Cast, initializing the Gabor parameters of every FAT block ($f_{s,\tau}$, $\gamma_{s,\tau}$, $\boldsymbol{\mu}_{s,\tau}$, $\mathbf{w}_{s,\tau}$, $b_{s,\tau}$, and the family multiplier λ_s) from the Stage-1 values instead of from hand-selected ones; all remaining weights, including the lifting, the projection, and the MLP stream, are initialized afresh. Both stages follow the schedule of Sec. 4.2, and the Stage-1 parameters are taken from its best-validation checkpoint, reached between epochs 15 and 35 depending on the dataset. The wavelet family and level are kept as in Table 16.

Results. Table 9 compares the two-stage recipe with the tuned model of Table 1. CSI-M changes by between -0.15% (Shanghai) and -2.88% (CIKM), and the two-stage model still exceeds the strongest baseline of Table 1 in CSI-M on every dataset. The recipe trades a small amount of aggregate skill for skill at the highest threshold: CSI-H improves on MeteoNet ($+0.58\%$), Shanghai ($+0.84\%$), and CIKM ($+6.26\%$) and decreases only on SEVIR (-3.81%). On CIKM, the reported two-stage run is the best of five seeds, whose CSI-M spans 0.3150–0.3313; the 2.9% gap to the tuned model is therefore of the same order as CIKM's seed-to-seed spread. The advantage of DAWN-Cast over the baselines therefore does not depend on the per-dataset Gabor search.

G Additional Related Work

Recurrent models. ConvLSTM Shi et al. (2015) replaces the fully connected transitions of an LSTM with convolutions, so the hidden state keeps its spatial layout. TrajGRU Shi et al. (2017) observes that convolutional recurrence is location-invariant while rotation and scaling are not, and learns location-variant connections between hidden states. MAU Chang et al. (2021) widens the temporal receptive field of the recurrent unit by attending over several past states and fusing the result with the current one. PhyDNet Guen & Thome (2020) splits its latent state into a PhyCell branch,

whose dynamics are constrained to approximate partial differential operators, and a ConvLSTM branch that models the residual.

Recurrence-free models. SimVP Gao et al. (2022a) folds time into channels and shows that a purely convolutional encoder, translator and decoder is competitive with recurrent models. Earthformer Gao et al. (2022b) makes space-time attention tractable by restricting it to local cuboids within a hierarchical encoder–decoder. exPreCast Song et al. (2026) combines local spatiotemporal attention with a texture-preserving decoder and a temporal extractor for variable forecast horizons, targeting fine-scale structure in extreme events. EarthFarseer Wu et al. (2024) runs a local convolutional branch in parallel with a global Fourier-domain branch, so one layer mixes information at several spatial frequencies.

Generative models. DiffCast Yu et al. (2024) predicts the predictable part of the evolution with a deterministic backbone and models the residual with a diffusion process.

Design summary. Table 10 contrasts the design choices that distinguish DAWN-Cast from the two decomposition-based baselines.

Table 10: Design comparison of DAWN-Cast with the decomposition-based nowcasters AlphaPre Lin et al. (2025) and WADEPre Liu et al. (2026). ✓: present; ×: absent; –: not applicable (no scale decomposition).

Method	Wavelet decomposition	Separate temporal operator per scale	Cross-scale interaction before reconstruction	Spatial mixing in temporal operator	Adaptive Gabor temporal mapping
AlphaPre	×	×	–	✓	×
WADEPre	✓	×	✓	✓	×
DAWN-Cast (ours)	✓	✓	×	×	✓

DAWN-Cast and WADEPre. Both models decompose with a fixed discrete wavelet transform, give the approximation band and the detail bands separate networks, reconstruct with the inverse transform, and refine the result. They differ in what is decomposed, in whether the per-band networks see space, and in how detail levels are parameterized (Table 11).

H Additional Zero shot results

For zero-shot transfer from SEVIR to CIKM refer Table 12, all models are trained exclusively on SEVIR with $T_{out} = 20$ and evaluated on the CIKM test set without target-domain data, fine-tuning, or parameter updates. Each model receives the 5 observed CIKM input frames and generates its native 20-step forecast; to match the CIKM evaluation horizon, only the first 10 predicted frames are scored, while the remaining predictions are discarded. We use each model's native best checkpoint selected on SEVIR, ensuring that no model-specific adaptation to CIKM is performed. This protocol therefore evaluates direct cross-dataset generalization under a consistent 5-to-10 forecasting setting.

I Generalization and Adaptation Details

Table 13 provide additional analysis of target-domain adaptation. Increasing the adaptation surface to 44.57% of the parameters by updating srst block instead of FAT blocks refer Table 5. Table 14 quantifies the computational cost of adapting the SEVIR-trained model compared with training a target-domain model from scratch. Using only 20% of the target data reduces wall-clock training time by $5.3\times$ on MeteoNet and $4.6\times$ on Shanghai, while using 50% yields $2.2\times$ and $2.1\times$ speed-ups, respectively. The advantage largely disappears when the full target dataset is used, with with adaptation taking about 0.9× the time of training from scratch (1.1× speed-up). Indicating that the compute savings primarily arise from processing fewer target samples rather than from freezing parameters. Notably, on Shanghai, 20% target-data adaptation achieves higher CSI-M than full training from scratch (0.4639 vs. 0.4525) while requiring only 2.52 hours.

Table 11: Component-level comparison of WADEPre Liu et al. (2026) and DAWN-Cast.

	WADEPre	**DAWN-Cast**
Input to the DWT	Raw radar sequence	Lifted features (per-frame convolutional lifting)
Wavelet, levels	bior2.4, $J = 3$	db6/db6/db6/db4; $J = 2/1/3/2$ (SEVIR/MeteoNet/Shanghai/CIKM)
Approximation band	A-Net: 2D conv encoder with time as channels, 3D-conv temporal injector, dilated ResNet blocks (spatiotemporal)	FAT block: Gabor and MLP streams along time, fused by a $1\times1\times1$ conv; no spatial receptive field
Detail bands	One D-Net for all levels: shared temporal MLP, feature pyramid with level-specific stages, per-level IDR conv blocks	One FAT block per level, no parameters shared across levels; no spatial receptive field
Orientations (LH/HL/HH)	Stacked as channels	Stacked as channels; share temporal weights within a level
Cross-level interaction before reconstruction	Bidirectional, through the feature pyramid	None
Reconstruction	Three IDWTs: Y_A (predicted approx., repeated last details), Y_D (repeated last approx., predicted details), Y_{AD}	Two IDWTs: fused outputs, and Gabor outputs only
After reconstruction	Refiner (kinematic-coupling and drift-correction conv streams) added to Y_{AD}	Two RR blocks of SRST (Fourier ‖ depthwise conv, GN, SiLU, channel MLP) and a spectral layer; Gabor reconstruction added afterwards
Training objective	Multi-scale curriculum (ZNCC on approx., level-weighted MSE on details, mixed loss)	FACL loss
Lead-time prediction	$T_{\text{out}} = T_{\text{in}}$ (*not flexible*)	$T_{\text{out}} \neq T_{\text{in}}$(*flexible*)

Table 12: **Zero-shot transfer from SEVIR to CIKM.** All models are trained on SEVIR (T_{out}=20) and evaluated on the CIKM test set (T_{out}=10) without target-domain data or parameter updates: each model receives the 5 CIKM input frames, forecasts its native 20 steps, and only the first 10 steps are scored. All models use their native best checkpoints; bold marks the best value of each column.

Model	Params	CSI-M↑	HSS↑	MSE↓
ExPreCast	32.00M	0.2929	0.3877	48.08
WADEPre	30.42M	0.2883	0.3795	49.03
TrajGRU	11.91M	0.2850	0.3801	47.16
EarthFormer	8.65M	0.2796	0.3750	51.37
AlphaPre	89.03M	0.2774	0.3651	47.36
EarthFarseer	96.42M	0.2643	0.3558	51.80
SimVP	11.06M	0.2642	0.3520	52.89
MAU	7.63M	0.2619	0.3497	50.47
ConvLSTM	0.89M	0.2597	0.3441	56.54
PhyDNet	3.09M	0.2495	0.3350	50.58
DiffCast (PhyDNet)	49.36M	0.2315	0.3102	60.48
DAWN-Cast (ours)	59.54M	**0.3023**	**0.3967**	**40.45**

J Computational Cost

Table 15 reports parameters, FLOPs, inference time, and peak memory for all models under a single protocol: one sequence (batch size 1, 128×128, T_{in}=5, T_{out}=20), FP32 precision, a single NVIDIA RTX A6000, and the same forward path used for evaluation. FLOPs are counted with `fvcore` and reported as $2\times$ multiply–accumulates; inference time is the median of 30 timed forward passes after 10 warm-up passes; peak memory is the maximum CUDA allocation during one forward pass.

Table 13: **Wider-surface adaptation of the SEVIR DAWN-Cast to a new radar domain.** Lifting, projection, and `srst_block2` (the second block of the SRST residual refinement stage) are unfrozen; the temporal path (Gabor+MLP+fusion) stays *frozen*, unlike the 0.71%-surface variant (**26,540,353 of 59,541,089 parameters, 44.57%**). Each run is 50 epochs, $\mathrm{lr} = 10^{-4}$, batch 4. *Data* is the fraction of the target training set used (validation/test always full). Δ is relative to DAWN-Cast trained from scratch on that dataset; $+$ means adaptation outperforms it. Only 20% and 50% data points were run for this surface; no 100%-data run was performed.

Dataset	Adaptation	Data	CSI-M↑	ΔCSI	HSS↑	ΔHSS	CSI-pool4↑	CSI-pool16↑	MSE↓
MeteoNet	Full-shot	100 %	0.4529	-	0.5890	-	0.5159	0.6132	9.31
	Zero-shot	0%	0.3929	-13.24%	0.5206	-11.61%	0.4310	0.4963	9.76
	Fine-tuned (44.57% params)	20%	0.4415	-2.52%	0.5773	-1.99%	0.4964	0.5892	9.64
	Fine-tuned (44.57% params)	50%	0.4490	-0.87%	0.5871	-0.32%	0.5033	0.5688	9.21
Shanghai	Full-shot	100 %	0.4525	-	0.5918	-	0.5113	0.5842	25.33
	Zero-shot	0%	0.4422	-2.27%	0.5786	-2.23%	0.4925	0.5727	23.40
	Fine-tuned (44.57% params)	20%	0.4624	+2.19%	0.6010	+1.55%	0.5279	0.6251	25.11
	Fine-tuned (44.57% params)	50%	0.4659	+2.96%	0.6029	+1.88%	0.5163	0.5924	23.73

Table 14: **Cost of adapting a SEVIR model versus training a target model from scratch.** Wall-clock is the measured sum of per-epoch times from the training logs (single RTX A6000, batch 4, 50 epochs in every row). Speed-up is from-scratch time divided by adaptation time. The saving comes from the reduced target training set, not from freezing: a frozen parameter still needs its activations for backpropagation, so the 100% row costs almost as much as training from scratch. Note the Shanghai 20% row: 4.6× less compute and a better model than training from scratch.

Dataset	Regime	Trainable	Data	Time(sec)/epoch	Wall-clock (h)	Speed-up	CSI-M↑
MeteoNet	from scratch	59,541,089	100%	3654	50.75	1.0×	0.4529
	adaptation (0.71%)	423,009	20%	689	9.57	5.3×	0.4386
	adaptation (0.71%)	423,009	50%	1629	22.62	2.2×	0.4476
	adaptation (0.71%)	423,009	100%	3226	44.81	1.1×	0.4527
Shanghai	from scratch	59,541,089	100%	826	11.47	1.0×	0.4525
	adaptation (0.71%)	423,009	20%	181	2.52	4.6×	0.4639
	adaptation (0.71%)	423,009	50%	402	5.59	2.1×	0.4699
	adaptation (0.71%)	423,009	100%	775	10.77	1.1×	0.4703

Table 15: **Computational cost.** Parameters are as reported in Table 4. FLOPs, inference time, and peak memory are measured for one sequence (batch size 1) at 128×128, T_{in}=5, T_{out}=20, using FP32 on a single RTX A6000. Each model is profiled through the same forward path used for evaluation. GFLOPs are computed with `fvcore` and reported as $2\times$ multiply-accumulates. Inference time is the median over 30 timed forwards after 10 warm-up runs; peak memory is the maximum CUDA allocation during one forward.

Model	Params	GFLOPs	Inference (ms)	Peak memory (GB)
ConvLSTM	0.89M	364.4	20.8	0.14
PhyDNet	3.09M	153.2	49.9	0.03
MAU	7.63M	179.8	101.8	0.11
EarthFormer	8.65M	70.5	15.0	0.29
SimVP	11.06M	411.4	66.6	0.18
TrajGRU	11.91M	313.2	237.8	0.27
WADEPre	30.42M	3262.6	184.9	0.77
ExPreCast	32.00M	26.6	13.3	0.17
DiffCast (PhyDNet)	49.36M	–[‡]	12.3 s	0.34
AlphaPre	89.03M	3113.3[†]	153.0	1.98
EarthFarseer	96.42M	1506.6[†]	135.2	1.51
DAWN-Cast (ours)	**59.54M**	**2089.5**[†]	**200.3**	**1.92**

[†] Excludes FFTs in the spectral blocks because `fvcore` does not provide formulas for `fft_rfft2`/`fft_irfft2`; these values are therefore lower bounds. Element-wise activations are uncounted for all models. [‡] DiffCast uses a 250-step diffusion sampler, whose computation graph is not traceable by `fvcore`. Its cost is dominated by 4×250 denoiser evaluations, which are reflected in the measured inference time.

Where the parameters are. Of DAWN-Cast's 59.54M parameters, the lifting, projection, and wavelet temporal path together hold only 0.42M (0.71%, Table 5); the remainder lies in the refinement stage, whose layers operate on $D = T_{out} \cdot c$ channels. The component that carries our contribution is therefore small.

Parameter count does not explain the gains. Three comparisons control for size. (i) A shared-FAT variant with matched parameters (59.51M vs. 59.54M) transfers clearly worse than DAWN-Cast (Table 4). (ii) The two largest baselines, AlphaPre (89.03M) and EarthFarseer (96.42M), are below DAWN-Cast on every metric in Table 1. (iii) Removing WGTM changes the parameter count by only 0.27% ($59.54\text{M} \rightarrow 59.38\text{M}$ on SEVIR; at most 1.03% on any of the four datasets) (Table 15) but lowers CSI-M on both SEVIR and CIKM (Table 2).

Inference. DAWN-Cast forecasts all T_{out} frames in a single forward pass.

K Additional Hyperparameter details

Table 16 lists the search space and the configuration selected per dataset by validation CSI-M. The search covers the wavelet family and level J, the depthwise kernel size k of the SRST local stream, the number of spectral groups N and GroupNorm groups ρ of the SRST block, the family multipliers λ_s of Eq. 1 (f_ℓ for the low-frequency family and f_h for the high-frequency families), the initial Gabor bandwidths (β_ℓ, β_h), and the initial Gabor weight scales (W_ℓ, W_h).

What is and is not fixed by the search. The multipliers and initial bandwidths set only the starting point of the Gabor stream: the per-frame frequencies $f_{s,\tau}$, bandwidths $\gamma_{s,\tau}$, weights, and prototypes remain learnable in every family (Sec. 3.1.2). The search therefore does not impose an ordering between families, and none is assumed: $f_h > f_\ell$ is selected on SEVIR and CIKM, the two are nearly equal on MeteoNet, and $f_h < f_\ell$ on Shanghai. Because the phase $\lambda_s f_{s,\tau}(\mathbf{w}_{s,\tau}^\top \mathbf{x} + b_{s,\tau})$ depends on the magnitude of the wavelet coefficients, which differs between datasets, multiplier values are not comparable across datasets.

Dependence on the search. The two-stage recipe (Appendix F) initializes the Gabor stream from a first-stage model instead of searching these hyperparameters. It changes CSI-M by between -0.15% and -2.88% and still exceeds the strongest baseline of Table 1 on every dataset (on CIKM, the reported two-stage run is the best of five seeds).

Other settings. db6 is selected on all datasets except CIKM (db4), and the wavelet level ranges from $J = 1$ (MeteoNet) to $J = 3$ (Shanghai). The SRST settings are shared by SEVIR, MeteoNet, and Shanghai ($k = 3$, $N = 4$, $\rho \in \{3, 4\}$); CIKM, whose domain is the smallest (101 km), selects a larger kernel ($k = 7$) and a single spectral group. On CIKM, a smaller initial weight scale for the high-frequency families ($W_h = 0.25$) was needed for stable early training.

Table 16: Hyperparameter search space and best values found per dataset. **Bold** indicates the selected value used in the final model.

Hyperparameter	Search Range	SEVIR	MeteoNet	Shanghai	CIKM
Wavelet	Haar, db4, db6	**db6**	**db6**	**db6**	**db4**
Wavelet-level (J)	1,2,3,4	**2**	**1**	**3**	**2**
Spatial Kernel K	{3, 7, 9}	**3**	**3**	**3**	**7**
SRST(Spectral Branch) N	{1, 2, 4, 8}	**4**	**4**	**4**	**1**
SRST (Spectral Branch)ρ	{1, 2, 3, 4, 8}	**4**	**4**	**3**	**1**
Low-Freq Multiplier f_ℓ	[0.10 - 30.00]	**0.1**	**1.09**	**1.09**	**22.74**
High-Freq Multiplier f_h	[0.10 - 100.00]	**4.0**	**1.12**	**0.14**	**95.56**
Bandwidth initialization β_ℓ	[0.05-100.00]	**0.17**	**0.0995**	**0.0995**	**95.56**
Bandwidth initialization β_h	[0.10 - 5.00]	**0.17**	**0.1643**	**0.1653**	**4.8193**
Weight initialization W_ℓ	$\{0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0\}$	**0.1**	**0.1**	**0.1**	**0.1**
Weight initializationW_h	$\{0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0\}$	**1.0**	**1.0**	**1.0**	**0.25**

L Statistical Analysis

Table 17 repeats DAWN-Cast training on CIKM with five seeds. The five-seed mean CSI-M (0.3358) is 1.6% below the seed-0 value in Table 1 but remains 3.0% above the strongest baseline (EarthFormer, 0.3260), and the mean CSI-40 (0.1598) exceeds EarthFormer's 0.1559, the only metric on which Table 1 ranks DAWN-Cast second. All baselines are single seed-0 runs, so margins in the main tables should be read against the seed standard deviation reported here.

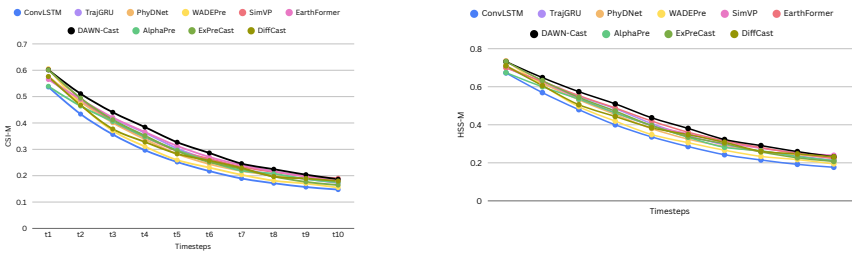
Table 17: **Statistical analysis of DAWN-Cast over five independent multiseed (0,1,2,3,4) runs on CIKM.** We report the mean $\pm$ standard deviation across runs.

Dataset	CSI-M ↑	CSI-40 ↑	HSS ↑	PSNR ↑
CIKM	0.3358 ± 0.0049	0.1598 ± 0.0072	0.4334 ± 0.0059	23.4655 ± 0.2374

M Lead-Time Analysis

Figure 8 reports CSI-M and HSS-M on CIKM separately for each of the 10 predicted frames (6-min steps, $+6$ to $+60$ min). Both are averaged over the CIKM thresholds (20, 30, 35, 40); Table 1 reports these curves averaged over lead times.

DAWN-Cast has the highest CSI-M and HSS-M from the second to the ninth predicted frame. Its margin is largest at intermediate lead times (t_3–t_6, $+18$ to $+36$ min), where the baselines separate most. At the first frame, PhyDNet and ExPreCast match or slightly exceed it, and by the last frame the strongest curves (DAWN-Cast, SimVP, and EarthFormer) converge. The aggregate gain on CIKM therefore comes mainly from intermediate lead times rather than from the final frame. At $+60$ min, the forecasts nevertheless differ visibly in structure (Fig. 3): CSI and HSS score thresholded pixels independently, so they do not reward a coherent storm shape over fragmented exceedances with similar pixel counts.



(a) CSI-M (b) HSS-M

Figure 8: **Lead-time performance on CIKM.** (a) CSI-M and (b) HSS-M at each predicted frame (t_1=+6 min to t_{10}=+60 min), averaged over the thresholds 20, 30, 35, and 40. DAWN-Cast is shown in black.

N Additional Qualitative Results

We show additional qualitative results comparing DAWN-Cast and 11 other baselines on all the 4 datasets.

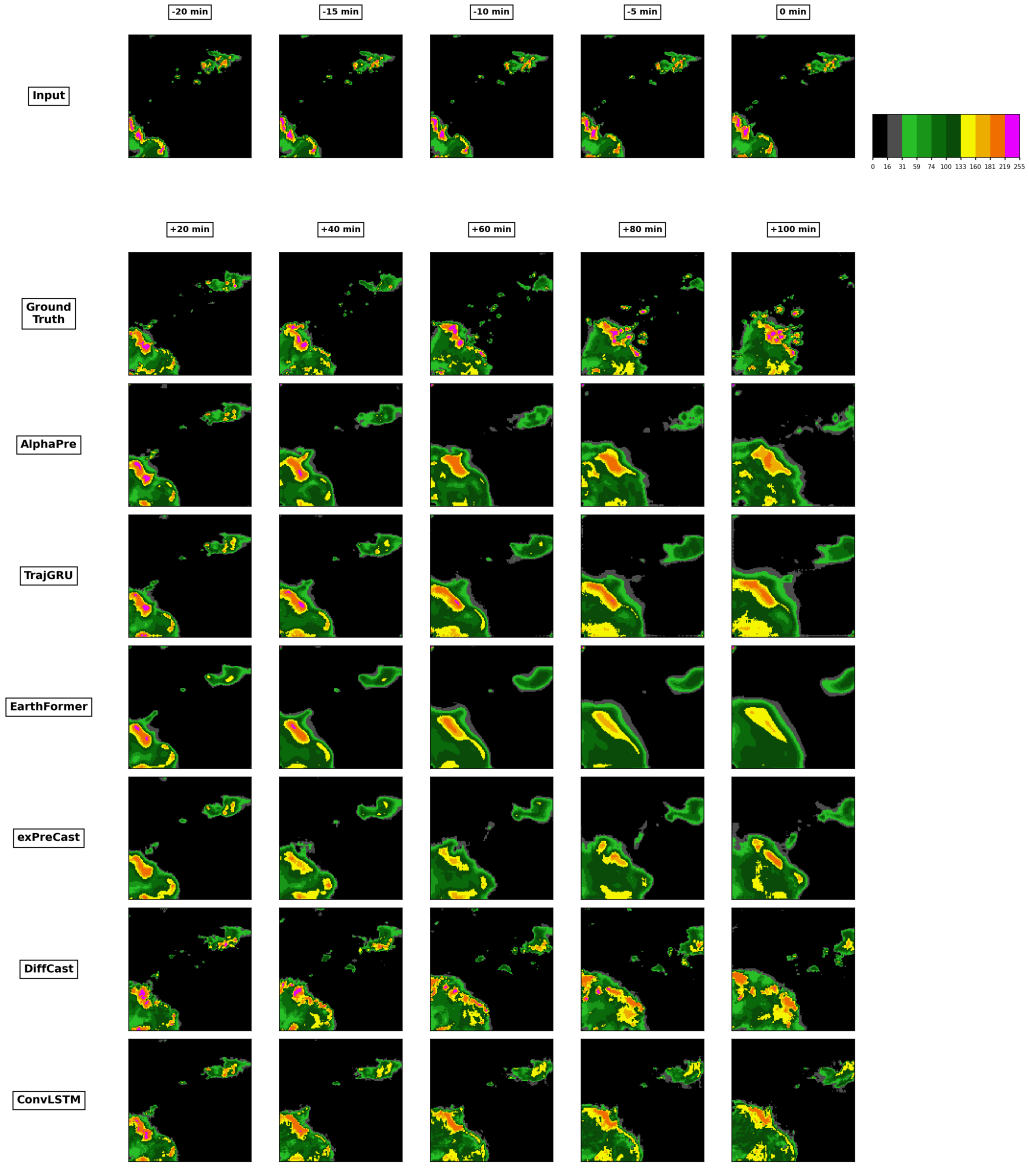


Figure 9: Qualitative comparision of SEVIR (part 1 of 2).

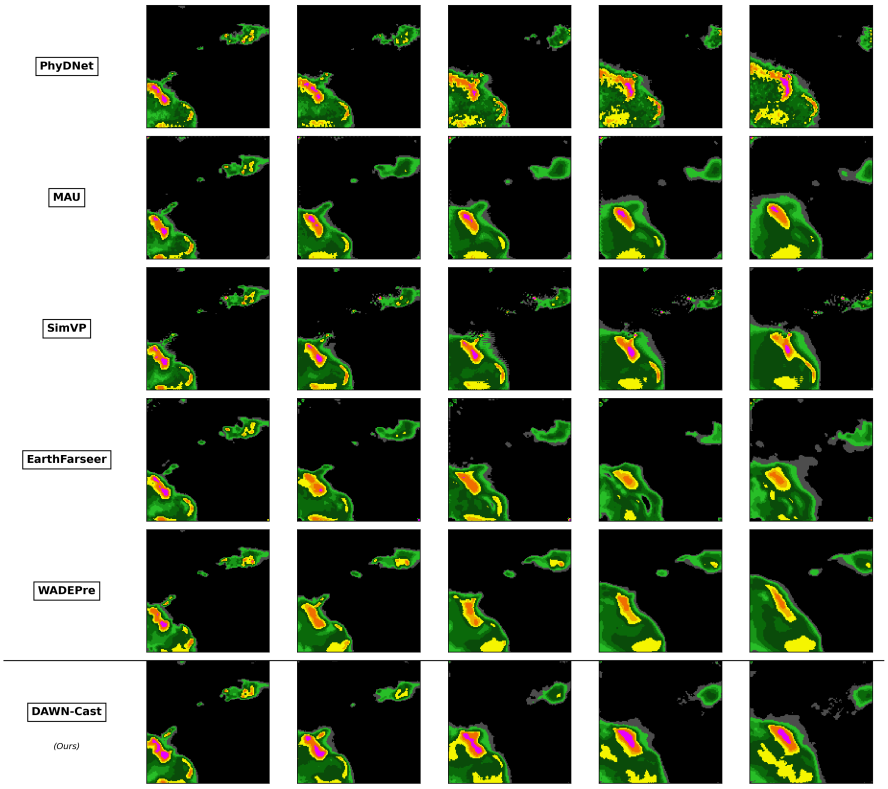


Figure 9: Qualitative comparision of SEVIR (part 2 of 2).

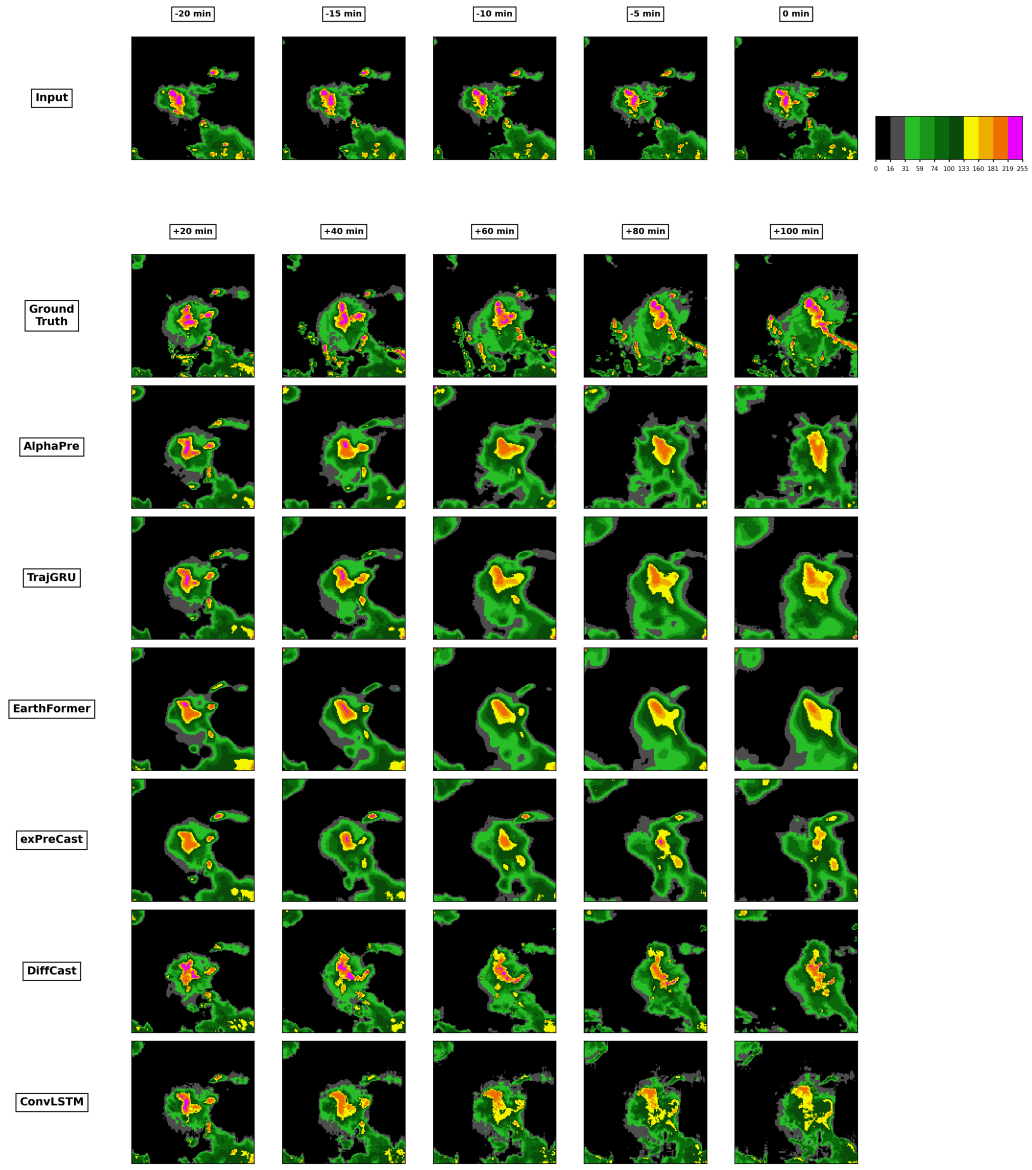


Figure 10: Qualitative comparision of SEVIR (part 1 of 2).

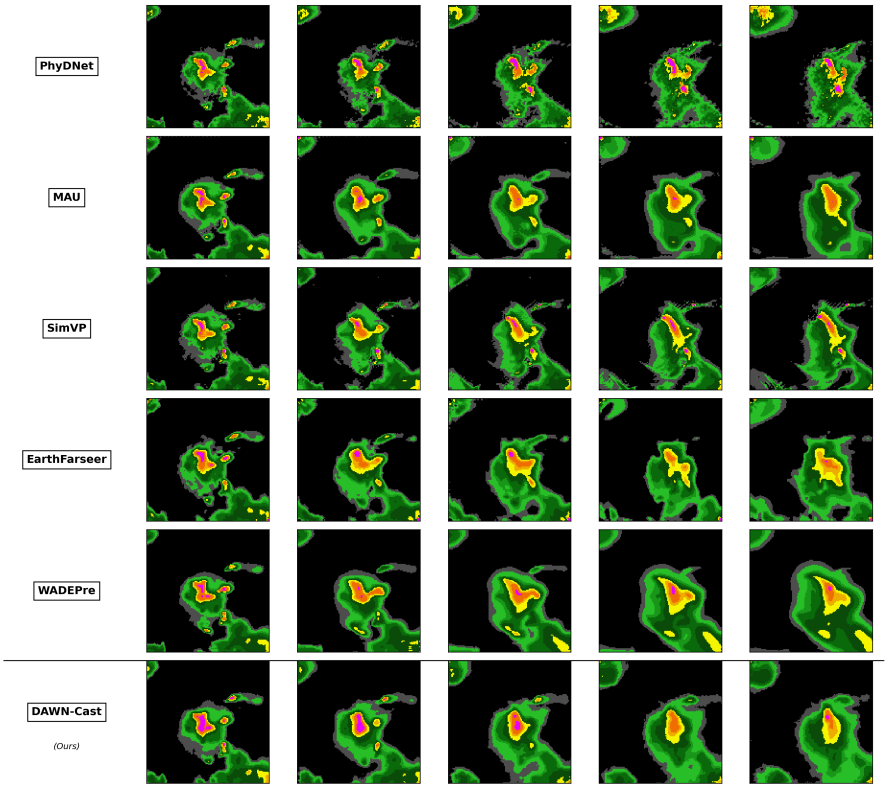


Figure 10: Qualitative comparision of SEVIR (part 2 of 2).

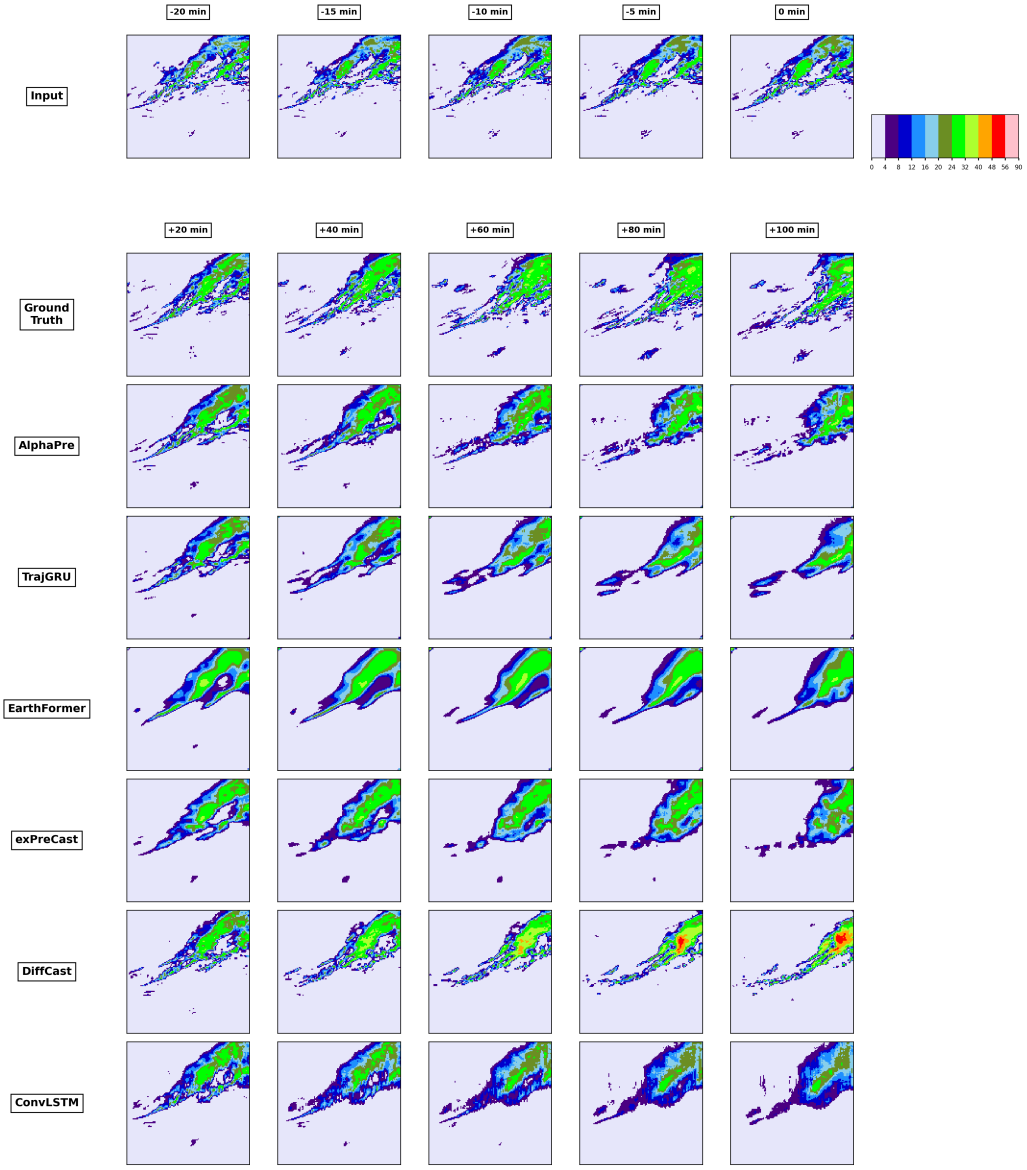


Figure 11: Qualitative comparision of Meteonet (part 1 of 2).

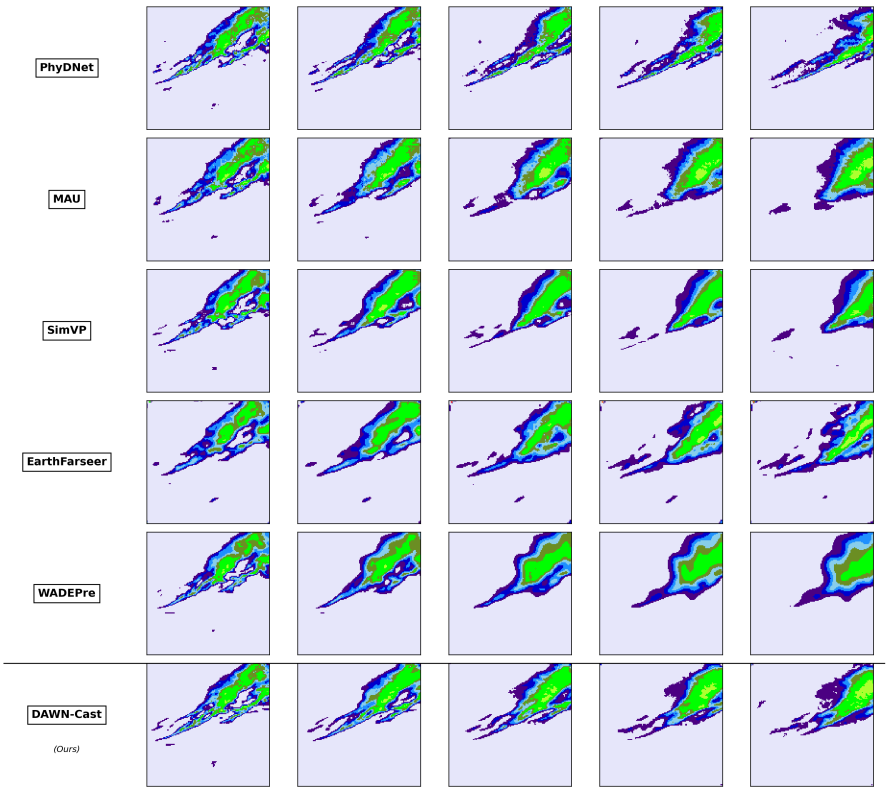


Figure 11: Qualitative comparision of Meteonet (part 2 of 2).

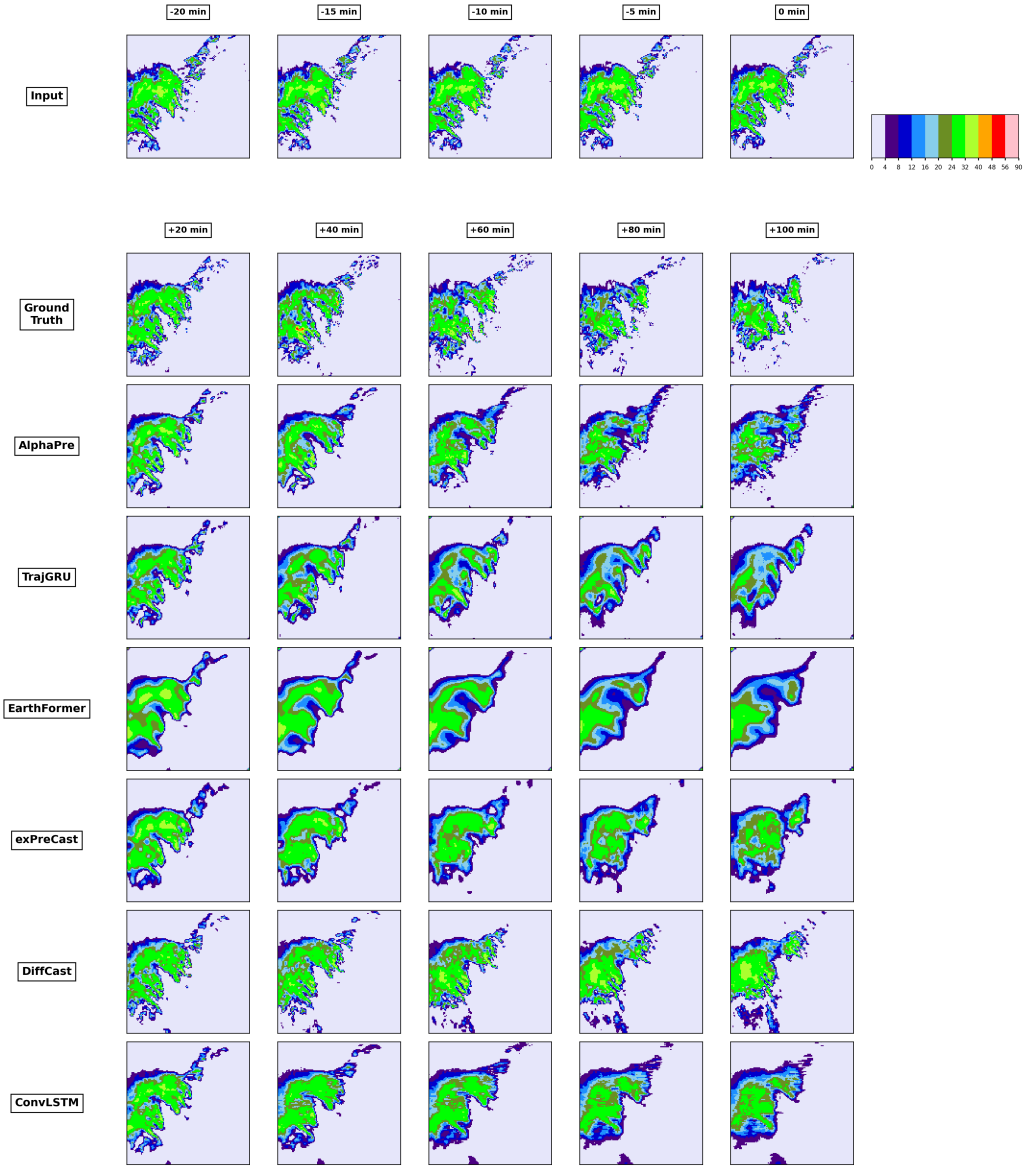


Figure 12: Qualitative comparision of Meteonet (part 1 of 2).

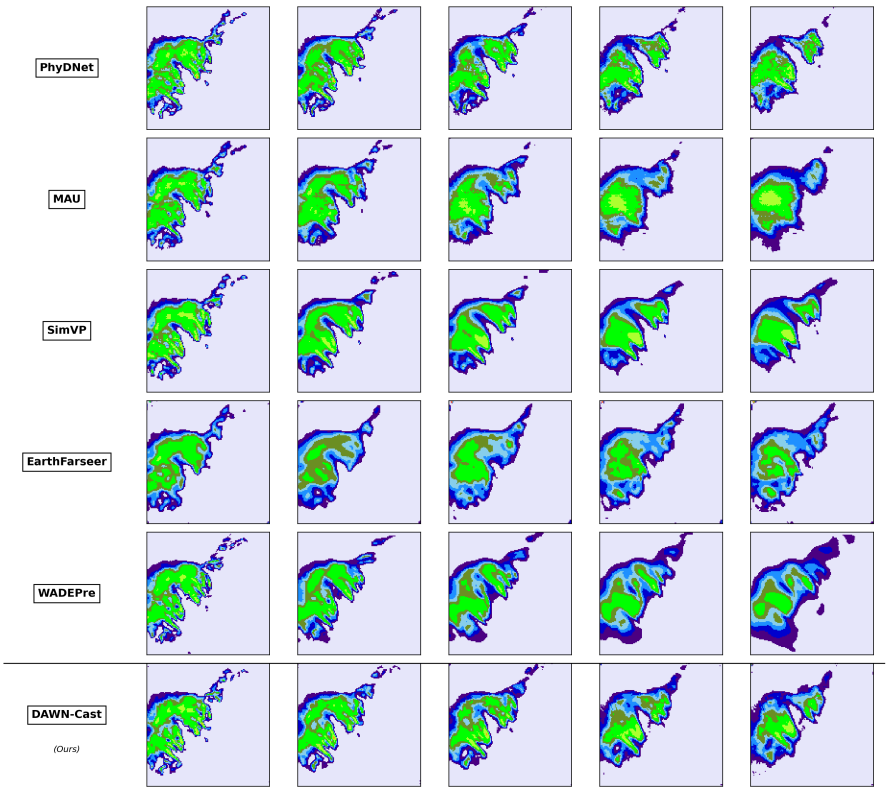


Figure 12: Qualitative comparision of Meteonet (part 2 of 2).

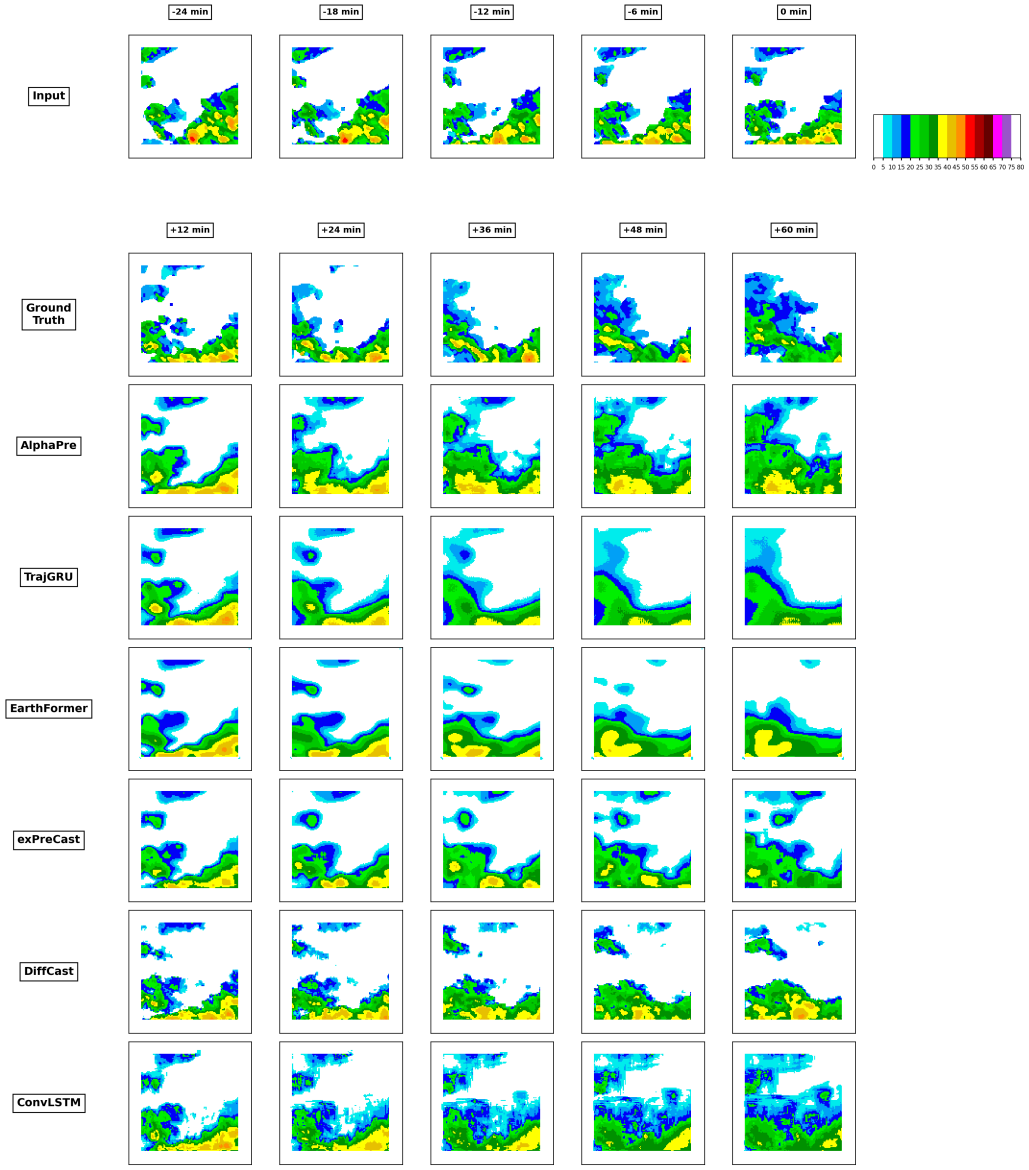


Figure 13: Qualitative comparision of CIKM (part 1 of 2).

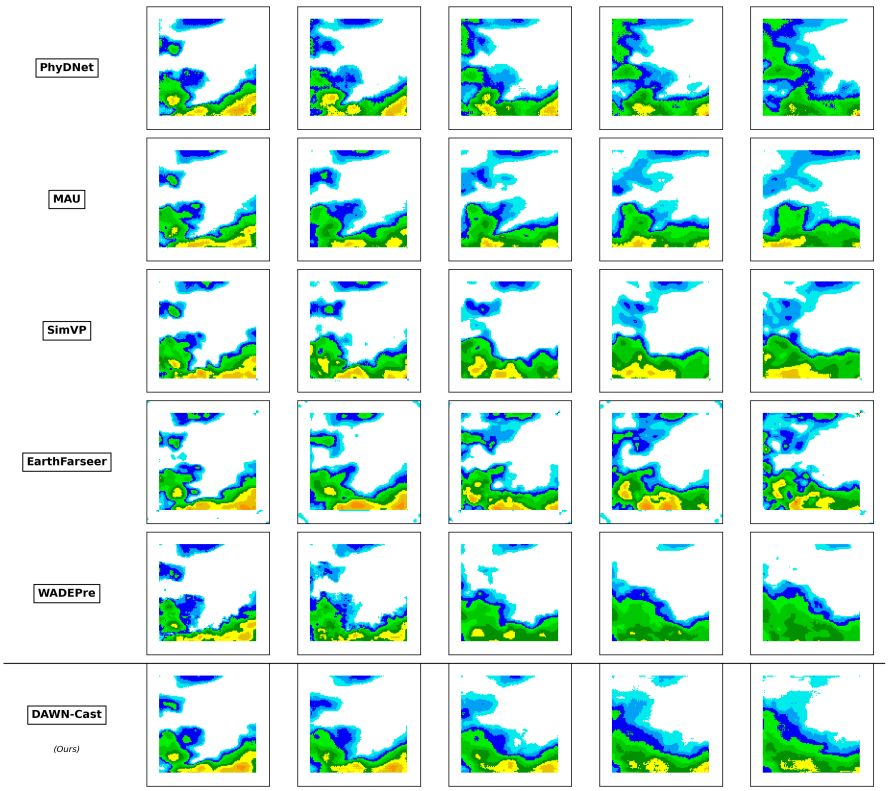


Figure 13: Qualitative comparision of CIKM (part 2 of 2).

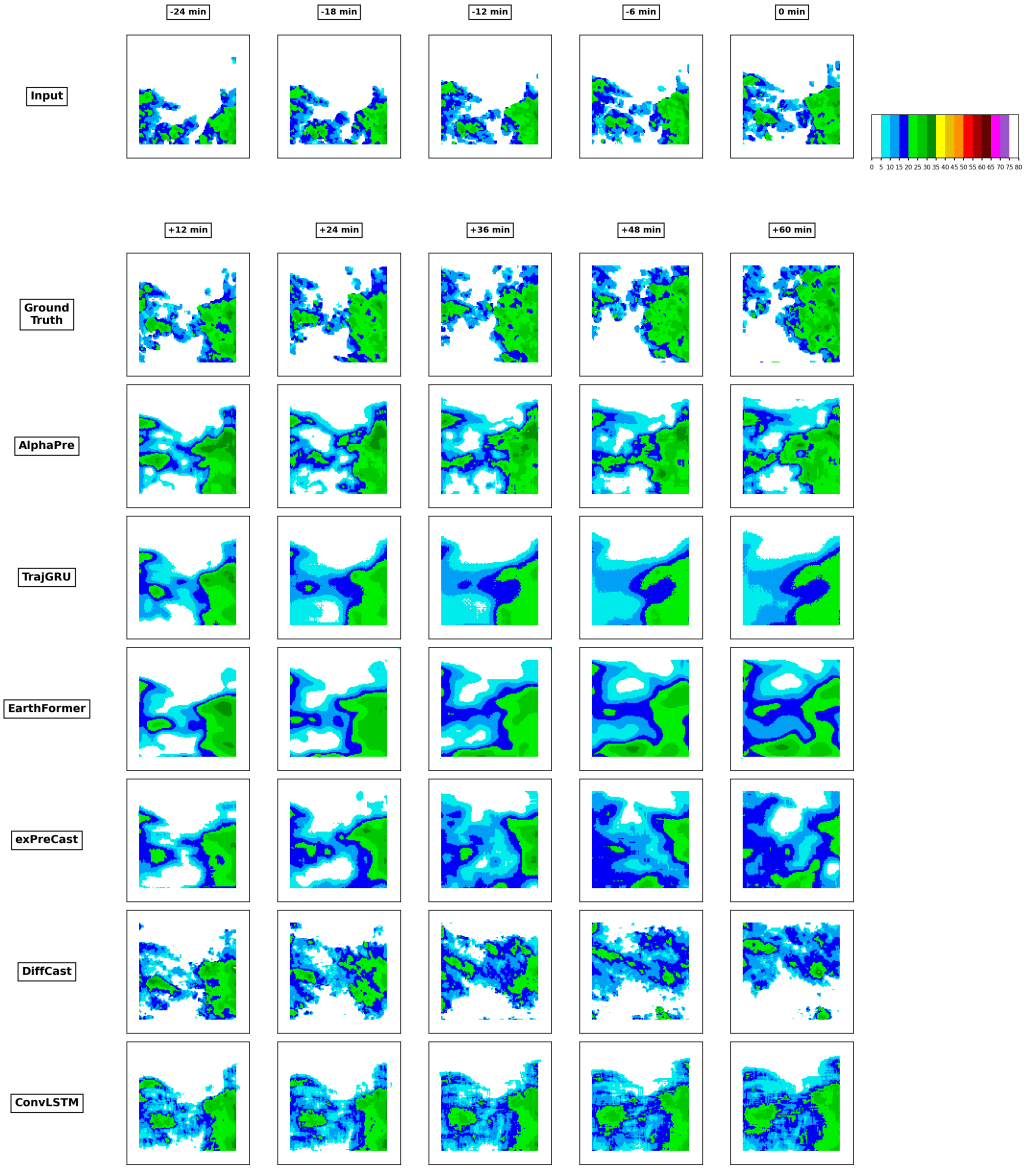


Figure 14: Qualitative comparision of CIKM (part 1 of 2).

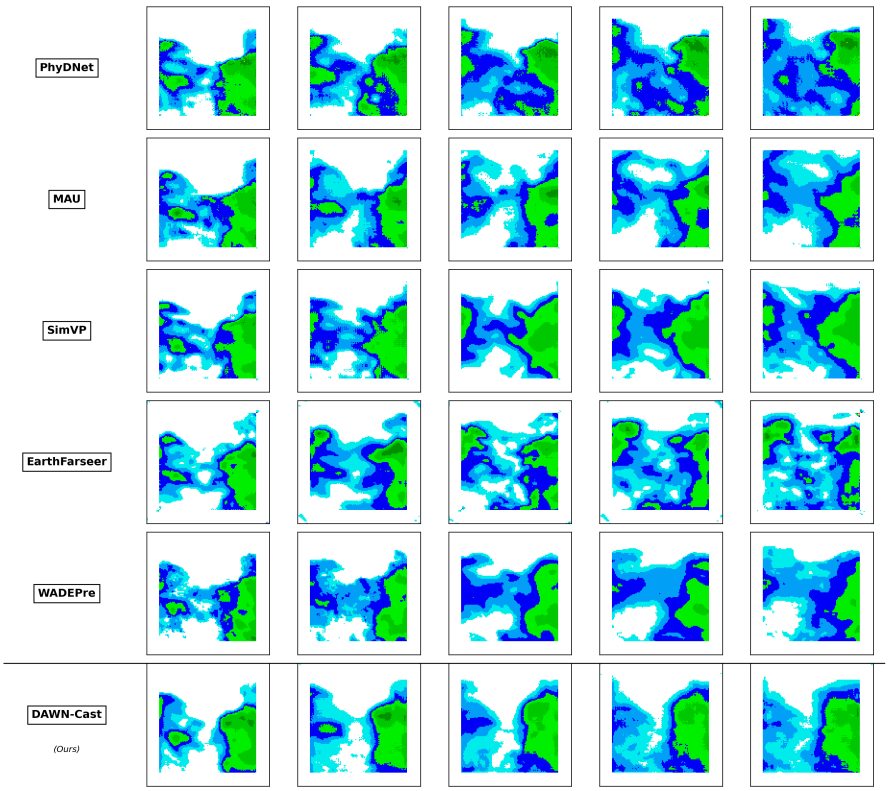


Figure 14: Qualitative comparision of CIKM (part 2 of 2).

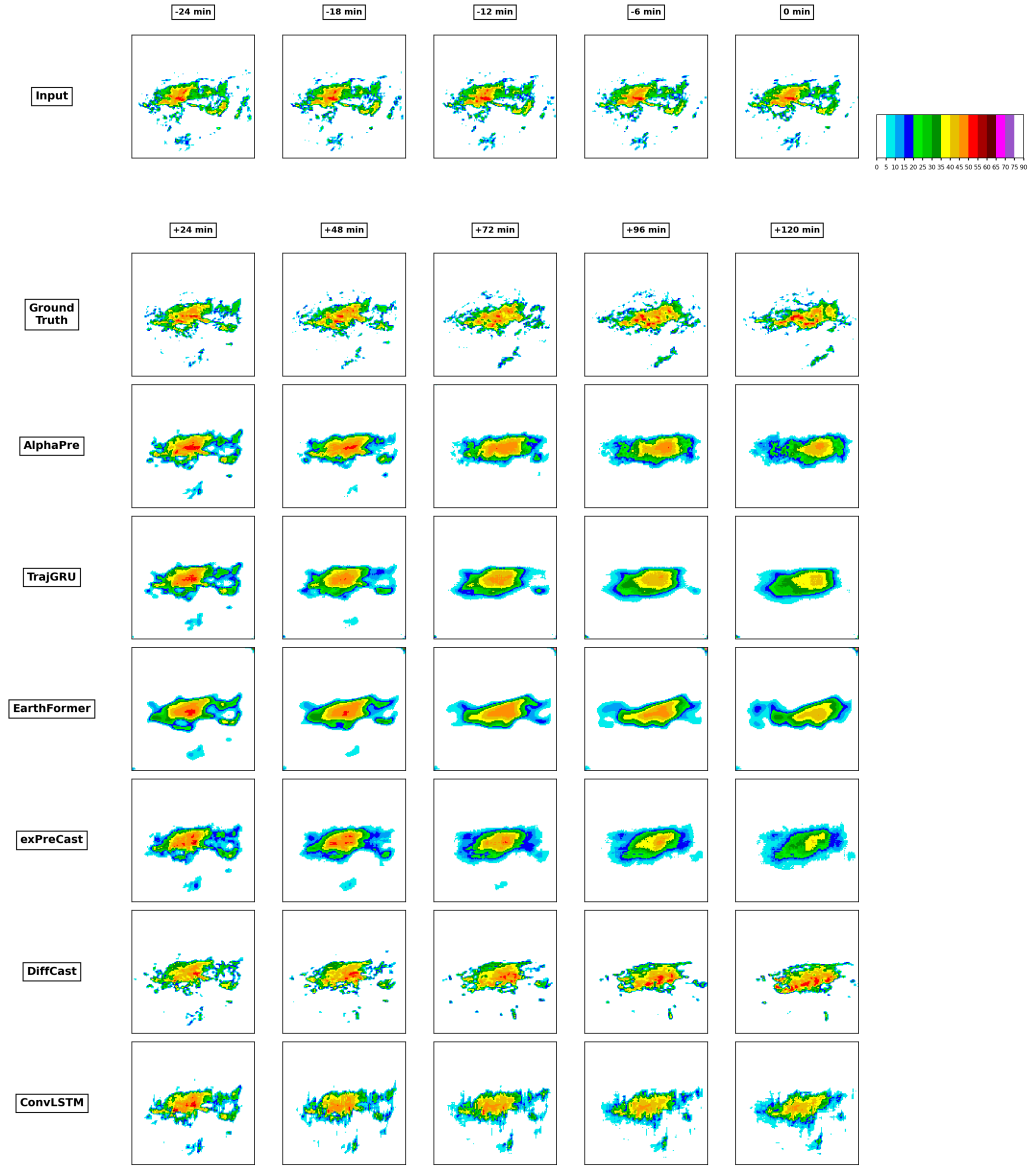


Figure 15: Qualitative comparision of Shanghai (part 1 of 2).

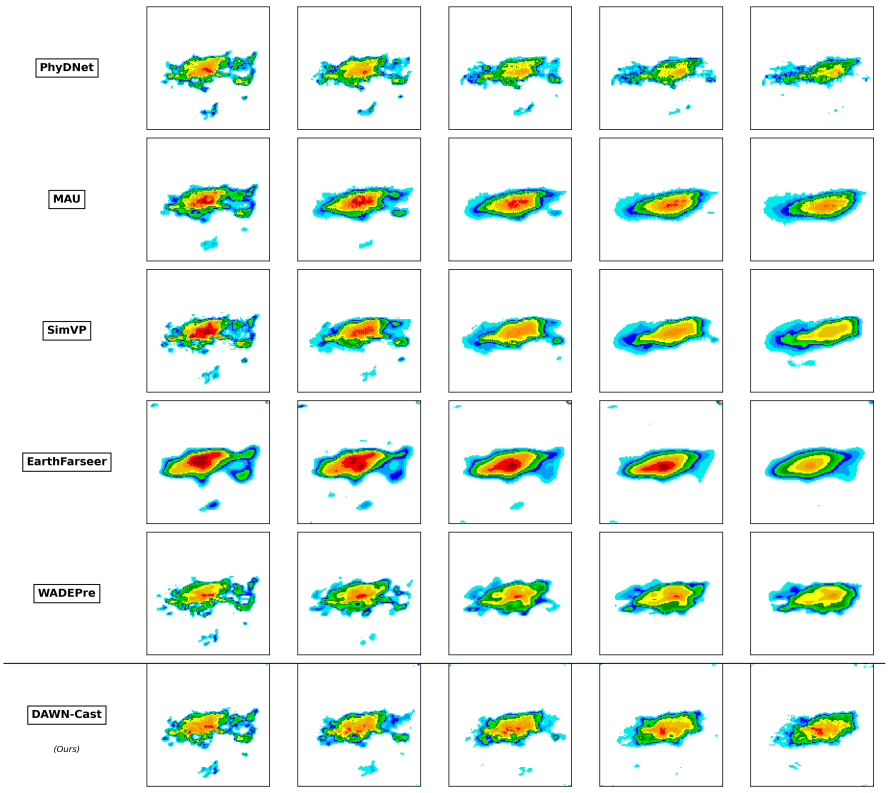


Figure 15: Qualitative comparision of Shanghai (part 2 of 2).

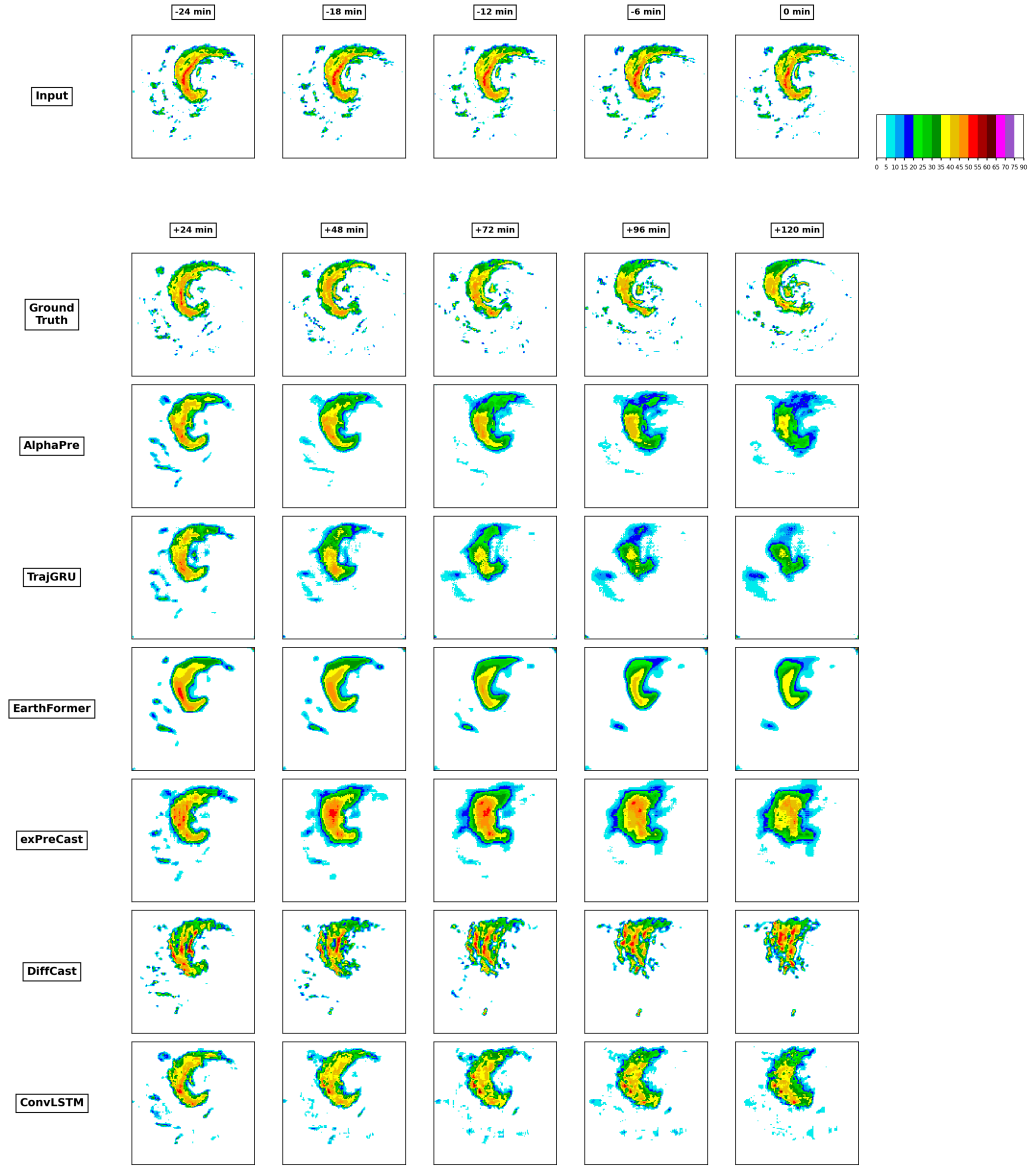


Figure 16: Qualitative comparision of Shanghai (part 1 of 2).

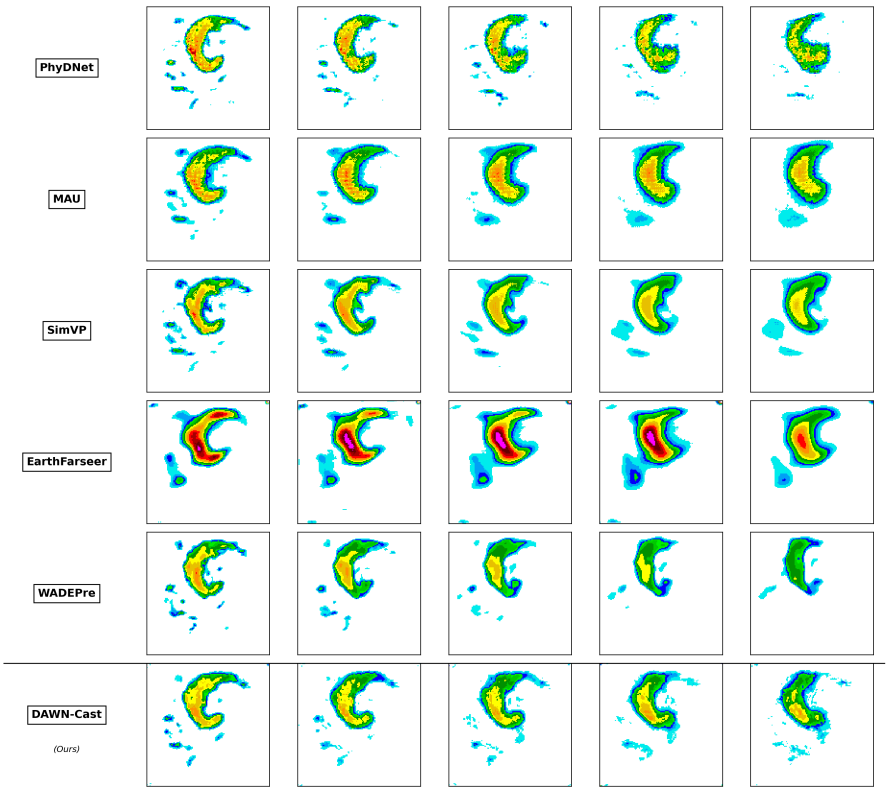


Figure 16: Qualitative comparision of Shanghai (part 2 of 2).