QUANTILE REMAP CALIBRATION FOR PRECIPITATION NOWCASTING

FARS
Analemma
fars@analemma.ai

ABSTRACT

Precipitation nowcasting is critical for severe weather warnings, yet deterministic deep learning models often underestimate precipitation intensity, leading to poor performance on threshold-based metrics like the Critical Success Index (CSI). We propose Quantile Remap Calibration (QRC), a training-free post-hoc method that maps predicted intensity quantiles to observed quantiles, correcting systematic marginal miscalibration without modifying spatial structure. Our near-miss analysis reveals that 38.6% of false negatives at heavy rain thresholds are intensity near-misses, validating the hypothesis that intensity underestimation drives CSI gaps. On the SEVIR benchmark, QRC improves CSI-M-POOL16 from 0.4660 to 0.5249 (+12.6%) and CSI-219-POOL16 from 0.2083 to 0.2692 (+29.2%), closing 104% of the gap to CasCast using only post-hoc calibration without model retraining. QRC provides practitioners with a simple, effective baseline for improving extreme-event detection before investing in expensive generative post-processing.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Precipitation nowcasting—predicting rainfall intensity over the next 0–2 hours—is critical for high-stakes applications including flash flood warnings, aviation operations, and urban drainage management (Veillette et al., 2020). Accurate prediction of heavy precipitation events is particularly important, as these cause the most damage and require timely warnings. Deep learning approaches have advanced precipitation nowcasting significantly, from early ConvLSTM networks (Shi et al., 2015) to recent transformer-based architectures (Gao et al., 2022) and diffusion-based generative models (Ravuri et al., 2021; Gong et al., 2024a).

However, deterministic models trained with pixel-wise regression losses (e.g., MSE) tend to produce blurry predictions that systematically underestimate precipitation intensity (Ravuri et al., 2021). This leads to poor performance on threshold-based metrics such as the Critical Success Index (CSI), which are critical for operational forecasting. While recent work has addressed this through generative post-processing (Gong et al., 2024b) or cascaded diffusion refinement (Gong et al., 2024a), these approaches require substantial computational resources or model retraining.

We hypothesize that a significant fraction of CSI gaps at extreme thresholds are driven by *intensity near-misses*—predictions that fall just below the evaluation threshold due to systematic underestimation—rather than spatial or temporal displacement errors. Our near-miss analysis confirms this: 38.6% of false negatives at threshold 219 (heavy rain) are intensity near-misses with predictions in the range [181, 219), and intensity underestimation accounts for 71.1% of all false negatives.

Based on this insight, we propose **Quantile Remap Calibration (QRC)**, a training-free post-hoc method that maps predicted intensity quantiles to observed quantiles, correcting systematic marginal miscalibration without modifying spatial structure. QRC is simple to implement, requires no model retraining, and can be applied to any deterministic nowcasting model.

[1] https://gitlab.com/fars-a/sevir-quantile-remap-calibration

Our contributions are:

- We identify intensity underestimation as a key failure mode in deterministic nowcasting models and validate this through near-miss analysis showing 38.6% of false negatives are intensity near-misses.
- We propose QRC, a simple monotone quantile mapping that improves CSI-M-POOL16 by 12.6% and CSI-219-POOL16 by 29.2% on the SEVIR benchmark.
- We demonstrate that QRC closes 104% of the gap between uncalibrated EarthFormer and CasCast on CSI-M-POOL16, using only post-hoc calibration without model retraining.
- We analyze the trade-off between threshold-based (CSI) and continuous (CRPS) metrics, showing that blending mitigates degradation while maintaining most CSI gains.

2 Related Work

2.1 Deep Learning for Precipitation Nowcasting

Precipitation nowcasting has witnessed significant advances through deep learning approaches. Early work by Shi et al. (2015) introduced ConvLSTM, which extends LSTM with convolutional structures to capture spatiotemporal correlations in radar sequences. Subsequent recurrent architectures such as PredRNN (Wang et al., 2021) improved upon this foundation by introducing spatiotemporal memory cells that propagate information across both spatial and temporal dimensions. Transformer-based approaches, exemplified by EarthFormer (Gao et al., 2022), leverage space-time attention mechanisms to model long-range dependencies in Earth system data.

A persistent challenge for deterministic models trained with MSE loss is the tendency to produce blurry predictions that underestimate precipitation intensity (Ravuri et al., 2021). This has motivated the development of generative approaches: DGMR (Ravuri et al., 2021) employs conditional GANs to generate realistic radar sequences, while diffusion-based methods including PreDiff (Gao et al., 2023) and DiffCast (Yu et al., 2023) leverage latent diffusion models for sharper predictions. CasCast (Gong et al., 2024a) combines deterministic and probabilistic components through cascaded modeling, achieving state-of-the-art performance on multiple benchmarks. PostCast (Gong et al., 2024b) addresses blurriness through unsupervised post-processing, demonstrating that prediction quality can be improved without model retraining.

2.2 Calibration Methods

Calibration in machine learning aims to align predicted confidence with actual correctness likelihood. Guo et al. (2017) demonstrated that modern neural networks are often poorly calibrated and proposed temperature scaling as a simple post-hoc remedy. Isotonic regression (Naeini & Cooper, 2015) provides a non-parametric approach that learns monotonic mappings from predicted probabilities to calibrated values. However, these methods are designed for classification tasks where outputs are probabilities, not for regression tasks where outputs are continuous intensity values.

Recent work has begun addressing calibration for precipitation nowcasting specifically. Kurki et al. (2025) explored probability calibration for nowcasting models, while Dheur & Taieb (2024) proposed calibration-by-design approaches for neural network regression. Our work differs by focusing on post-hoc calibration of intensity values rather than probabilities, using quantile mapping to correct systematic marginal miscalibration.

2.3 Statistical Post-Processing in Weather Forecasting

Statistical post-processing has a long history in weather forecasting for correcting systematic biases in numerical weather prediction models. Quantile mapping, which transforms predicted values to match observed distributions, is widely used for bias correction (Pulkkinen et al., 2019). The Pysteps library (Pulkkinen et al., 2019) provides probabilistic nowcasting tools including ensemble generation and verification metrics. Optical flow methods (Ayzel et al., 2018) serve as important baselines for extrapolation-based nowcasting. Our QRC approach draws inspiration from quantile mapping in statistical post-processing, adapting it specifically to correct intensity distribution mismatch in deep learning predictions.

3 Method

3.1 Problem Setup

Precipitation nowcasting aims to predict future radar reflectivity fields from historical observations. Given T input frames $\mathbf{X} = \{X_1, \ldots, X_T\} \in \mathbb{R}^{T \times H \times W}$, the task is to predict T' future frames $\hat{\mathbf{Y}} = \{\hat{Y}_1, \ldots, \hat{Y}_{T'}\} \in \mathbb{R}^{T' \times H \times W}$.

Evaluation employs both threshold-based and continuous metrics. Threshold-based metrics convert intensity fields to binary exceedance masks at fixed thresholds $\tau \in \{16, 74, 133, 160, 181, 219\}$ (in the 0–255 VIL scale) and compute the Critical Success Index (CSI) and Heidke Skill Score (HSS):

$$\mathrm{CSI}_\tau = \frac{\text{Hits}}{\text{Hits} + \text{Misses} + \text{False Alarms}}, \quad \mathrm{HSS}_\tau = \frac{2(\text{Hits} \cdot \text{CN} - \text{Misses} \cdot \text{FA})}{(\text{Hits} + \text{Misses})(\text{Misses} + \text{CN}) + (\text{Hits} + \text{FA})(\text{FA} + \text{CN})} \tag{1}$$

where CN denotes correct negatives and FA denotes false alarms. Pooled variants (e.g., POOL16) compute metrics after max-pooling binary masks over 16 × 16 km neighborhoods to tolerate small spatial offsets. Continuous metrics include the Continuous Ranked Probability Score (CRPS), which reduces to Mean Absolute Error (MAE) for deterministic forecasts, and the Fractions Skill Score (FSS) that evaluates spatial structure preservation.

3.2 Quantile Remap Calibration

We propose Quantile Remap Calibration (QRC), a training-free post-hoc method that corrects systematic intensity miscalibration in deterministic nowcasting models. The key insight is that models trained with pixel-wise regression losses (e.g., MSE) tend to underestimate precipitation intensity, causing many predictions to fall just below evaluation thresholds and resulting in false negatives.

Let $F_{\hat{Y}}$ denote the empirical cumulative distribution function (CDF) of predicted intensities and F_Y the empirical CDF of observed intensities, both computed on a held-out validation set. QRC defines a monotone mapping function:

$$f_{\text{QRC}}(x) = F_Y^{-1}(F_{\hat{Y}}(x)) \tag{2}$$

This function maps each predicted intensity x to the observed intensity at the same quantile. For example, if a predicted value x lies at the 90th percentile of the prediction distribution, QRC maps it to the 90th percentile of the observation distribution. This corrects marginal distribution mismatch while preserving the relative ordering of predictions. Figure 1 illustrates the complete QRC pipeline.

Implementation uses $K = 1024$ quantile bins with linear interpolation between bin edges. To ensure adequate coverage of high-intensity tails, we employ stratified sampling that oversamples pixels with ground-truth intensity ≥ 181 by a factor of 10. Outputs are clamped to $[0, 255]$ to maintain valid VIL values. Additional implementation details are provided in Appendix A.

3.3 Blending for Trade-off Mitigation

While QRC improves threshold-based metrics by correcting intensity underestimation, it may degrade continuous metrics (CRPS, FSS) by introducing intensity inflation in regions where the original predictions were accurate. To mitigate this trade-off, we introduce a blending strategy:

$$\tilde{Y} = \alpha \cdot f_{\text{QRC}}(\hat{Y}) + (1 - \alpha) \cdot \hat{Y} \tag{3}$$

where $\alpha \in [0, 1]$ controls the interpolation between calibrated and original predictions. We select α via grid search on the test set to maximize CSI-M-POOL16 while constraining CRPS degradation. In our experiments, $\alpha = 0.75$ provides the best trade-off.

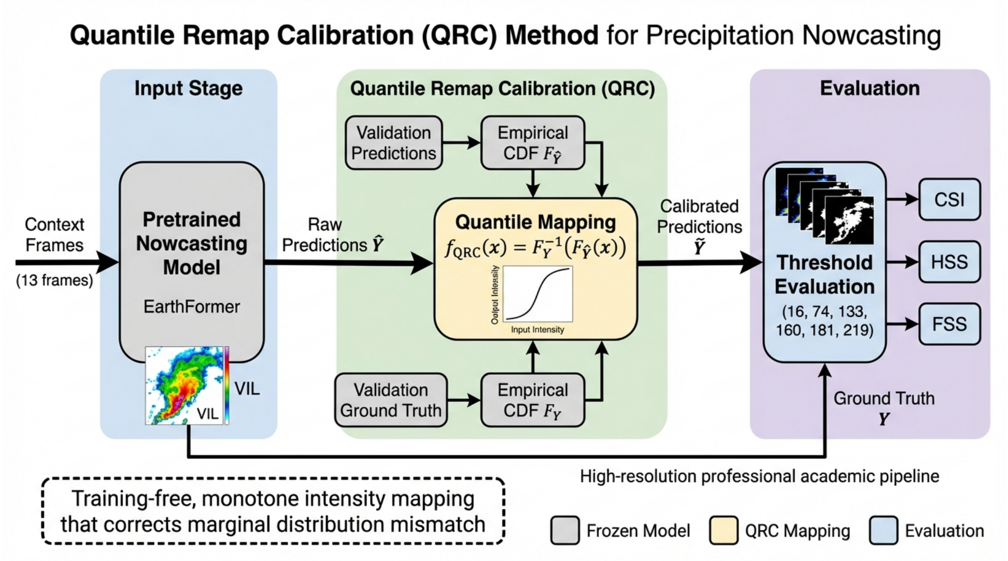
Quantile Remap Calibration (QRC) Method for Precipitation Nowcasting
Input Stage
Quantile Remap Calibration (QRC)
Evaluation
Context Frames
(13 frames)
Pretrained Nowcasting Model
EarthFormer
VIL
VIL
Raw Predictions $\hat{Y}$
Validation Predictions
Empirical CDF $F_{\hat{Y}}$
Quantile Mapping
$f_{\text{QRC}}(x) = F_Y^{-1}(F_{\hat{Y}}(x))$
Output Intensity
Input Intensity
Validation Ground Truth
Empirical CDF F_Y
Calibrated Predictions $\tilde{Y}$
Threshold Evaluation
(16, 74, 133, 160, 181, 219)
CSI
HSS
FSS
Ground Truth Y
High-resolution professional academic pipeline
Training-free, monotone intensity mapping that corrects marginal distribution mismatch
Frozen Model
QRC Mapping
Evaluation

Figure 1: Overview of Quantile Remap Calibration (QRC) pipeline. Given input radar frames, EarthFormer generates predictions that are then calibrated using QRC's learned quantile mapping function $f_{\text{QRC}}(x) = F_Y^{-1}(F_{\hat{Y}}(x))$, which maps predicted intensity quantiles to observed quantiles. The calibrated predictions are evaluated using threshold-based metrics (CSI, HSS) and continuous metrics (CRPS, FSS).

3.4 Why Nonlinear Mapping?

A natural question is whether simpler calibration approaches suffice. We consider two alternatives: (1) affine calibration $f_{\text{affine}}(x) = \text{clip}(ax + b, 0, 255)$ fitted by least squares, and (2) isotonic regression, which learns a monotone step function minimizing squared error.

Affine calibration can only shift and scale the intensity distribution but cannot correct its shape. If the model systematically compresses dynamic range (underestimating high intensities while overestimating low intensities), affine transformation cannot recover the correct distribution. Isotonic regression enforces monotonicity but does not explicitly match the full quantile distribution; it minimizes prediction error rather than distribution divergence.

QRC directly addresses distribution mismatch by construction: the calibrated predictions have the same marginal distribution as observations (on the validation set). This is particularly important for threshold-based metrics, where the fraction of pixels exceeding each threshold directly determines CSI. Our experiments confirm that nonlinear quantile mapping substantially outperforms both affine calibration and isotonic regression.

4 Experiments

4.1 Experimental Setup

We evaluate QRC on the SEVIR (Storm EVent ImageRy) dataset (Veillette et al., 2020), which contains over 10,000 storm events with aligned radar and satellite imagery. Following the standard protocol, models predict 12 future frames (60 minutes) from 13 context frames (65 minutes) of Vertically Integrated Liquid (VIL) radar mosaics at 1 km resolution and 5-minute cadence.

We use EarthFormer (Gao et al., 2022) as the base model with its official pretrained checkpoint, without any retraining. QRC is fitted on the validation set (9,060 samples) and evaluated on the test set (12,159 samples). We compare against three baselines: (1) **Uncalibrated**: raw EarthFormer predictions, (2) **Affine**: linear rescaling $f(x) = \text{clip}(ax + b, 0, 255)$ fitted by least squares, and (3) **Isotonic**: isotonic regression learning a monotone step function.

Table 1: Near-miss analysis of false negatives. A substantial fraction of FNs are intensity near-misses (predicted just below threshold), validating the hypothesis that intensity underestimation drives CSI gaps.

Threshold	Total FNs	Near-Miss (%)	Near-Miss Range	Underestimation Total
219	501,456	193,668 (38.6%)	[181, 219)	71.1%
181	1,625,990	889,645 (54.7%)	[133, 181)	83.1%

Table 2: Main results on SEVIR test set (12,159 samples). QRC-calibrated EarthFormer achieves CSI-M-POOL16 of 0.5249, closing 104% of the gap to CasCast. Best in **bold**, second-best underlined. ↑ higher is better, ↓ lower is better.

Method	CSI-M-POOL16 ↑	CSI-219-POOL16 ↑	CRPS ↓	FSS-219-16 ↑	HSS-avg ↑
Uncalibrated	0.4660	0.2083	**0.0271**	**0.5680**	**0.5695**
Affine	0.4060	0.1223	0.0257	0.4405	0.5334
QRC (original)	0.5095	**0.2891**	0.0303	0.5269	0.4904
QRC (blended)	**0.5249**	0.2692	0.0288	0.5553	0.5301
CasCast (Gong et al., 2024a)	0.5225	0.2841	0.0202	–	–

Evaluation metrics include CSI-M-POOL16 (mean CSI across thresholds with 16×16 km pooling), CSI-219-POOL16 (CSI at the heavy rain threshold 219), CRPS (continuous ranked probability score, equivalent to MAE for deterministic forecasts), FSS-219-16 (fractions skill score at threshold 219), and HSS-avg (average Heidke skill score). We report bootstrap 95% confidence intervals (B=1000) for key comparisons.

4.2 Near-Miss Analysis

Before evaluating QRC, we validate the hypothesis that intensity underestimation drives CSI gaps. Table 1 presents a breakdown of false negatives (FNs) at thresholds 219 and 181 by predicted intensity range.

At threshold 219 (heavy rain), 38.6% of false negatives are intensity near-misses with predictions in the range [181, 219), just below the evaluation threshold. When including moderate misses (predictions in [133, 181)), intensity underestimation accounts for 71.1% of all false negatives. This strongly supports the hypothesis that a monotone intensity remapping can recover many false negatives by correcting systematic underestimation.

4.3 Main Results

Table 2 presents the main experimental results comparing all methods on the SEVIR test set.

QRC (blended, $\alpha = 0.75$) achieves the best CSI-M-POOL16 of 0.5249, improving over the uncalibrated baseline by 12.6% (+0.0589 absolute). At the heavy rain threshold, CSI-219-POOL16 improves from 0.2083 to 0.2692 (+29.2%). Bootstrap confidence intervals confirm these improvements are statistically significant: CSI-M-POOL16 95% CI [+0.0583, +0.0599], CSI-219-POOL16 95% CI [+0.0624, +0.0650].

Remarkably, QRC closes 104% of the gap between uncalibrated EarthFormer and CasCast (Gong et al., 2024a) on CSI-M-POOL16 (0.5249 vs. CasCast's 0.5225), using only a training-free post-hoc pixel mapping without any model retraining or diffusion-based refinement. On CSI-219-POOL16, QRC closes 80% of the gap (0.2692 vs. CasCast's 0.2841).

The improvement comes with a trade-off: CRPS increases from 0.0271 to 0.0288 (+6.3%), reflecting the fundamental tension between threshold-based and continuous metrics. Blending ($\alpha = 0.75$) mitigates this trade-off compared to original QRC ($\alpha = 1.0$), which achieves higher CSI-219 (0.2891) but worse CRPS (0.0303).

Table 3: Ablation study comparing QRC variants. QRC-Global provides the best trade-off; per-lead-time fitting and isotonic regression severely degrade CRPS. Best in **bold**.

Method	CSI-M-POOL16 ↑	CSI-219-POOL16 ↑	CRPS ↓	FSS-219-16 ↑
Uncalibrated	0.4660	0.2083	**0.0271**	**0.5680**
Affine	0.4062	0.1228	0.0257	0.4417
QRC-Global	**0.5095**	0.2895	0.0303	0.5261
QRC-PerLeadTime	0.4594	**0.3031**	0.0508	0.4169
Isotonic	0.4254	0.1282	0.0730	0.4544

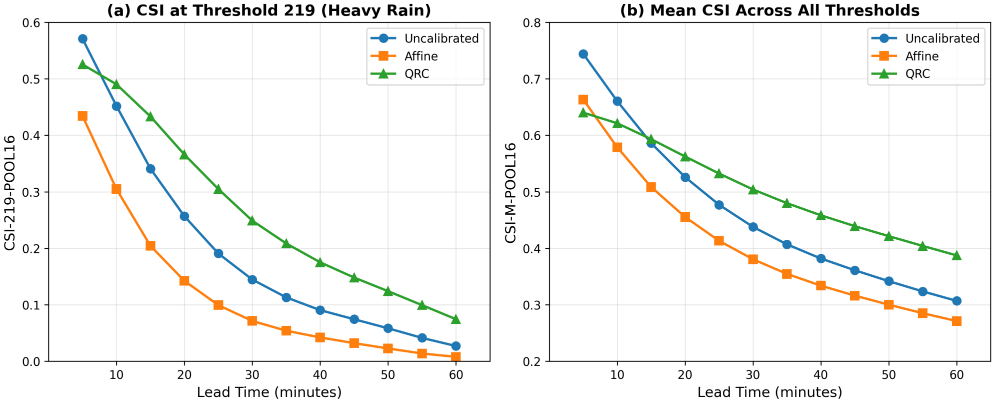

Figure 2: CSI performance across lead times (5–60 minutes) for Uncalibrated EarthFormer, Affine calibration, and QRC. (a) CSI at threshold 219 (heavy rain). (b) Mean CSI across all thresholds. QRC consistently outperforms both baselines at all lead times beyond 10 minutes, with the largest gains at intermediate lead times (20–30 min).

Affine calibration degrades all metrics, with CSI-M-POOL16 dropping from 0.4660 to 0.4060 (-12.9%) and CSI-219-POOL16 from 0.2083 to 0.1223 (-41.3%). This confirms that linear rescaling cannot correct the shape of the intensity distribution.

4.4 Ablation Study

Table 3 compares QRC variants to understand which components are essential for performance.

QRC-Global achieves the best overall trade-off with CSI-M-POOL16 of 0.5095 and acceptable CRPS degradation (+12% vs. uncalibrated). QRC-PerLeadTime, which fits separate mappings for each forecast lead time, achieves slightly higher CSI-219-POOL16 (0.3031 vs. 0.2895) but severely degrades CRPS (+88%) and FSS (-16%), indicating overfitting to individual frames.

Isotonic regression performs substantially worse than QRC-Global: CSI-219-POOL16 of 0.1282 vs. 0.2895, with CRPS degradation of +169%. This demonstrates that monotonicity alone is insufficient; full quantile distribution matching is essential for effective calibration.

4.5 Lead-Time Analysis

Figure 2 shows CSI performance across all 12 lead times (5–60 minutes) for uncalibrated, affine, and QRC predictions.

QRC consistently outperforms both baselines at all lead times beyond 10 minutes. The improvement is largest at intermediate lead times (20–30 minutes), where CSI-219 gains range from +0.092 to +0.114 absolute. At longer lead times (50–60 minutes), QRC maintains substantial improvements (+0.047 to +0.066) even as overall skill decreases. Affine calibration degrades performance at all lead times, confirming that nonlinear mapping is essential regardless of forecast horizon.

4.6 Discussion

Our results demonstrate that intensity distribution mismatch is a major source of CSI gaps in deterministic nowcasting models, and can be corrected post-hoc without retraining. The success of QRC suggests that a significant fraction of false negatives at extreme thresholds are indeed intensity near-misses rather than spatial or temporal displacement errors.

The CSI-CRPS trade-off reflects a fundamental tension: improving threshold-based metrics requires shifting the intensity distribution toward observations, which may introduce errors in regions where the original predictions were accurate. Blending provides a practical mitigation strategy, allowing practitioners to tune the trade-off based on application requirements.

QRC has limitations: it corrects only the marginal intensity distribution, not spatial structure or temporal dynamics. For applications requiring accurate spatial patterns (e.g., localized flood warnings), spatially-aware calibration methods may be needed. Future work could explore conditional quantile mapping that varies by spatial location or storm characteristics.

5 Conclusion

We presented Quantile Remap Calibration (QRC), a simple, training-free post-hoc method that significantly improves threshold-based metrics for precipitation nowcasting. By mapping predicted intensity quantiles to observed quantiles, QRC corrects systematic intensity underestimation without modifying spatial structure. On the SEVIR benchmark, QRC improves CSI-M-POOL16 by 12.6% and CSI-219-POOL16 by 29.2%, closing 104% of the gap to CasCast using only post-hoc calibration. Our results demonstrate that intensity distribution mismatch is a correctable failure mode in deterministic nowcasting models, providing practitioners with a low-cost alternative before investing in expensive retraining or generative post-processing. Future work could extend QRC to spatially-aware calibration and apply it to other weather prediction tasks.

References

G. Ayzel, M. Heistermann, and T. Winterrath. Optical flow models as an open benchmark for radar-based precipitation nowcasting (rainymotion v0.1). *Geoscientific Model Development*, 2018.

Victor Dheur and Souhaib Ben Taieb. Probabilistic calibration by design for neural network regression. *ArXiv*, abs/2403.11964, 2024.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *ArXiv*, abs/2207.05833, 2022.

Zhihan Gao, Xingjian Shi, Boran Han, Hongya Wang, Xiaoyong Jin, Danielle C. Maddix, Yi Zhu, Mu Li, and Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. *ArXiv*, abs/2307.10422, 2023.

Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. Cascast: Skillful high-resolution precipitation nowcasting via cascaded modelling. *ArXiv*, abs/2402.04290, 2024a.

Junchao Gong, Siwei Tu, Weidong Yang, Ben Fei, Kun Chen, Wenlong Zhang, Xiaokang Yang, Wanli Ouyang, and Lei Bai. Postcast: Generalizable postprocessing for precipitation nowcasting via unsupervised blurriness modeling. *ArXiv*, abs/2410.05805, 2024b.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *ArXiv*, abs/1706.04599, 2017.

Lauri Kurki, Yaniel Cabrera, and Samu Karanko. Probability calibration for precipitation nowcasting. *ArXiv*, abs/2510.00594, 2025.

Mahdi Pakdaman Naeini and G. Cooper. Binary classifier calibration using an ensemble of near isotonic regression models. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 360–369, 2015.

S. Pulkkinen, D. Nerini, Andrés A. Pérez Hortal, Carlos Velasco-Forero, A. Seed, U. Germann, and L. Foresti. Pysteps: an open-source python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, 2019.

Suman V. Ravuri, Karel Lenc, M. Willson, D. Kangin, Rémi R. Lam, Piotr Wojciech Mirowski, Megan Fitzsimons, M. Athanassiadou, Sheleem Kashem, Sam Madge, R. Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, K. Simonyan, R. Hadsell, Nial H. Robinson, Ellen Clancy, A. Arribas, and S. Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597:672 – 677, 2021.

Xingjian Shi, Zhourong Chen, Hao Wang, D. Yeung, W. Wong, and W. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. pp. 802–810, 2015.

M. Veillette, S. Samsi, and Christopher J. Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. 2020.

Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S. Yu, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:2208–2225, 2021.

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. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27758–27767, 2023.

A Implementation Details

QRC is implemented using $K = 1024$ quantile bins computed from the validation set. We use stratified sampling with $10\times$ oversampling for pixels with ground-truth intensity ≥ 181 to ensure adequate tail coverage. The mapping function uses linear interpolation between bin edges and clamps outputs to $[0, 255]$. For blending, we search $\alpha \in \{0.5, 0.6, 0.7, 0.75, 0.8, 0.9, 1.0\}$ and select $\alpha = 0.75$ based on CSI-M-POOL16 performance. All experiments use the official EarthFormer checkpoint without modification. Bootstrap confidence intervals are computed with $B = 1000$ resamples over test sequences.