Geophysical Research Letters

RESEARCH LETTER
10.1029/2025GL119442

Enhancing Machine Learning Models for Nowcasting and Short-Term Forecasting of Precipitation With a Novel Probability-Matching Loss Function

Yuan Cao[1], Shuaiyi Li[2], Jie Feng[2,3,4], Lei Chen[1], Hao Li[5], and Yijun Zhang[2,4]

[1]Shanghai Central Meteorological Observatory, Shanghai, China, [2]Department of Atmospheric and Oceanic Sciences and Institute of Atmospheric Sciences, Fudan University, Shanghai, China, [3]Key Laboratory of High Impact Weather (special), China Meteorological Administration, Changsha, Hunan, China, [4]Shanghai Key Laboratory of Ocean-Land-Atmosphere Boundary Dynamics and Climate Change, Shanghai, China, [5]Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai, China

Abstract Machine learning (ML) precipitation forecasting models typically employ a mean squared error (MSE) loss function for optimization. However, due to the well-known "double penalty" effect, MSE-based losses often lead to overly smoothed prediction fields and a systematic underestimation of the frequency of heavy rainfall. To address this limitation, we propose a novel probability-matching (PM) based loss for ML precipitation nowcasting and short-term forecasting, comparing its performance with other classical losses. Comprehensive skill evaluation demonstrates that PM-based loss offers relatively more balanced and consistent performance across metrics, particularly its lower forecast bias from light to heavy rainfall intensities. Spectral power analysis further indicates that PM-based loss better preserves small-scale precipitation variability throughout the forecast period. Additionally, it results in a forecast frequency distribution of precipitation that more closely aligns with the observed distribution. These findings indicate consistent improvements in predictive skill and reliability of ML precipitation forecasts when trained with the PM-based loss.

Plain Language Summary Accurately predicting heavy rainfall is still a major challenge for machine learning (ML) precipitation forecasting models. One of the key reasons is that the conventional loss functions in the training procedure of ML models tend to make precipitation forecasts too blurry and hardly predict the heavy rainfall. Thus, we develop a probability-matching (PM) based loss function to improve the model's forecasting ability. Our results show that PM-based loss delivers relatively more balanced performance. Unlike conventional loss functions, it does not systematically predict too little heavy rain events or too many light rain events. Forecasts using PM-based loss also keep much more realistic small-scale details as the prediction evolves over time, better showing how storms actually grow and move. Furthermore, the total amount of predicted rainfall at different intensities aligns more closely with observed rainfall. Overall, the PM-based loss consistently improves both the accuracy and reliability of ML models. This method could also be valuable for predicting other intermittent weather, such as fog or hail.

Key Points:

- A probability-matching (PM) loss function is proposed to mitigate rapid smoothing in machine learning-based precipitation forecasts
- PM-based loss function achieves relatively more balanced performance and lower bias across all rainfall intensities
- Forecasted precipitation frequency using PM-based loss matches observations more closely than MSE- and WMSE-based loss

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

J. Feng,
fengjiefj@fudan.edu.cn

Citation:

Cao, Y., Li, S., Feng, J., Chen, L., Li, H., & Zhang, Y. (2025). Enhancing machine learning models for nowcasting and short-term forecasting of precipitation with a novel probability-matching loss function. *Geophysical Research Letters*, *52*, e2025GL119442. https://doi.org/10.1029/2025GL119442

Received 12 SEP 2025
Accepted 19 NOV 2025

Author Contributions:

Conceptualization: Yuan Cao, Jie Feng
Formal analysis: Yuan Cao, Shuaiyi Li, Jie Feng, Lei Chen, Hao Li, Yijun Zhang
Investigation: Yuan Cao, Shuaiyi Li, Jie Feng, Lei Chen, Hao Li, Yijun Zhang
Methodology: Yuan Cao, Shuaiyi Li, Jie Feng
Validation: Yuan Cao, Shuaiyi Li, Jie Feng, Yijun Zhang
Visualization: Yuan Cao, Shuaiyi Li
Writing – original draft: Yuan Cao, Shuaiyi Li, Jie Feng

1. Introduction

Precipitation associated with meso-to convective-scale weather systems is typically characterized by high intensity and rapid evolution. The forecasting of such precipitation is generally classified into nowcasting (0–2 hr) and short-term forecasting (2–12 hr) (Ravuri et al., 2021; Shi et al., 2017; Zhang et al., 2023). These short-duration yet intense precipitation events can lead to severe sequences, such as flash flooding and landslides. As a result, accurate nowcasting and short-term forecasting are crucial for effective risk mitigation (Groenemeijer et al., 2017; Toth et al., 2000). However, predicting these events remains highly challenging due to their strong association with localized, small-scale convective storms, which are inherently short-lived and evolve rapidly (Brajard et al., 2021; Leutbecher et al., 2017).

In recent years, artificial intelligence (AI)-based techniques have been increasingly applied to meteorological forecasting, particularly in nowcasting and short-term prediction of precipitation (McGovern et al., 2023). These AI-based models for short-range precipitation forecasting build upon the foundational principles of extrapolation methods, leveraging spatio-temporally continuous observational data from previous time steps to predict future

Writing – review & editing: Jie Feng, Lei Chen, Hao Li, Yijun Zhang

conditions. By incorporating advanced machine learning (ML) techniques—such as ConvLSTM (Shi et al., 2015), MetNet (Espeholt et al., 2022; Sønderby et al., 2020), and EarthFormer (Gao et al., 2022)—these models are capable of capturing complex, nonlinear relationships between input data and target predictions.

Despite the effectiveness of ML-based nowcasting models, a critical limitation is their tendency to gradually smooth and blur rainfall structures as lead time increases. This results in the underestimation of high-intensity precipitation and the overestimation of weak events (see Figure 1a). The primary cause of this issue lies in the optimization process of the neural network parameters, which typically involves minimizing a predefined loss function. Most models employ loss functions such as the mean squared error (MSE), which quantify the spatially aggregated differences between the predicted and reference fields at the grid-point level (e.g., Arcomano et al., 2020; Grönquist et al., 2021; Sattari et al., 2025). However, this formulation tends to encourage the model to suppress extreme values in order to reduce the overall error and mitigate the so-called "double penalty" effect (see Section S8 in Supporting Information S1 for details), especially for longer lead times.

To mitigate the underestimation of extreme precipitation caused by the smoothing effect in ML-based nowcasting models, various techniques have been developed. A common strategy involves modifying the loss function to better capture the intensity of heavy precipitation events (e.g., Hu et al., 2021; Kim et al., 2021; Xu et al., 2019), for example, through resampling (Nooten et al., 2023) and reweighting (Wang et al., 2022) techniques. These methods enable the model to focus on heavy precipitation scenarios by either increasing the proportion of heavy rainfall samples or assigning higher weights to intense precipitation pixels within the loss function. However, these strategies often result in a trade-off, as it tends to significantly increase bias and false alarm rates for intense precipitation events. This issue partly arises from the subjective and empirical tuning of sampling weights across precipitation grids with varying intensity categories, underscoring the need for more principled and adaptive weighting schemes. On the other side, recent years have witnessed rapid advances in probabilistic forecasting models based on generative ML methods. Unlike deterministic ML models, generative models—such as DGMR (Ravuri et al., 2021), NowcastNet (Zhang et al., 2023), STGM (Wang et al., 2023), and Prediff (Gao et al., 2023)—aim to align the distribution of predictions with that of observations at the data set level. As an approach orthogonal to deterministic models, generative models exhibit notable potential for enhancing the representation of extreme events via probabilistic sampling, yet they differ significantly from deterministic models in the optimization objectives of their loss functions. This study primarily focuses on loss functions for deterministic models.

In recent decades, ensemble forecasting has become essential in Numerical Weather Prediction (NWP) for assessing forecast uncertainty and providing probabilistic insights (Bauer et al., 2015; Bremnes, 2004). Although the ensemble mean (EM) reduces forecast error compared to a single run, it often smooths out precipitation extremes, limiting its effectiveness (Feng et al., 2020). To address this, Ebert (2001) proposed the probability-matching (PM) EM, which retains the EM's spatial structure while adjusting precipitation intensities to match the distribution from ensemble members. The concept of PM has been widely applied in meteorological forecasting and postprocessing, particularly for short-range hydrological and flood warnings. Early studies employed PM as a metric to evaluate probabilistic predictive skill (Germann et al., 2006; Lin et al., 2005). Building on this, Bowler et al. (2006) and subsequent studies (Pierce et al., 2012; Seed et al., 2013) developed the STEPS probabilistic precipitation forecasting framework, which integrates extrapolation-based nowcasting with downscaled NWP outputs under the PM constraint. These deterministic and PM-based probabilistic nowcasting methods have since been implemented in the open-source Pysteps library (Pulkkinen et al., 2019) and extensively validated across diverse meteorological events (Imhoff et al., 2020). However, in all these studies, the PM concept has primarily been used for verification and postprocessing rather than for optimizing model performance.

Inspired by the PM-based EM and relevant applications, we propose a modified loss function for ML forecasting models that augments the traditional MSE-based loss. The proposed loss function includes a PM constraint that quantifies the aggregated differences between sorted predicted and observed precipitation values, thereby aligning their distributions regardless of spatial location. This PM-based loss function is applied to ML models for precipitation nowcasting and short-term forecasting and is evaluated against models trained with MSE loss and reweighting strategies. Given the persistent smoothing effect associated with MSE-based training, this proposed PM-based loss offers a promising and broadly applicable solution for improving the representation of precipitation in ML-based meteorological forecasts.

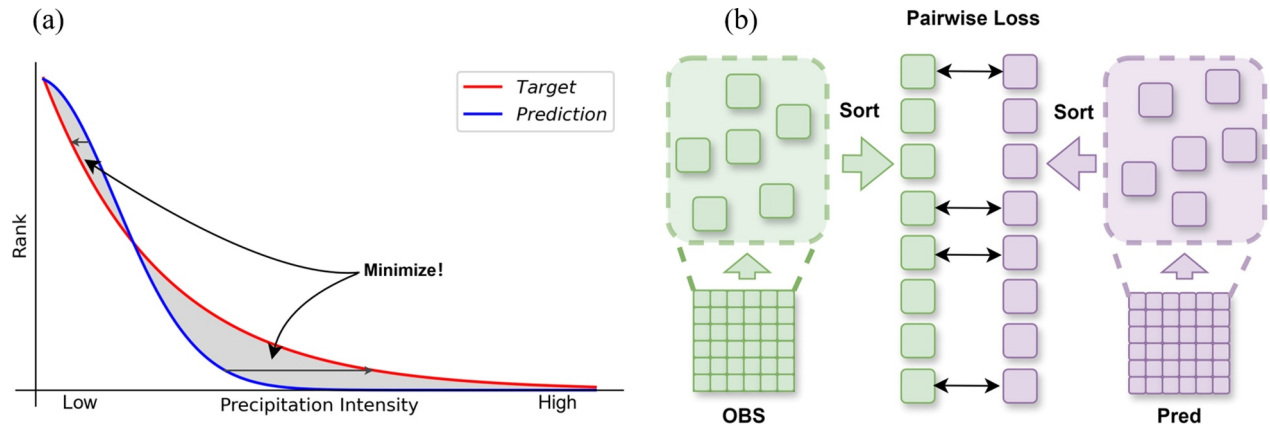

Figure 1. (a) The optimization objective of Probability-Matching (PM) based loss. (b) Graphical illustration of the calculation of PM loss between the target observation and the prediction.

2. Data for Training and Verification

In this study, we systematically evaluate the validity and effectiveness of the PM-based loss function by assessing its performance across various ML architectures in nowcasting and short-term prediction of precipitation. Given the distinct spatiotemporal characteristics of these forecasting tasks, we first introduce the data set used for training and evaluating the ML models in each scenario.

2.1. SEVIR for Nowcasting

For the nowcasting scenario, we utilize the Storm EVent ImageRy (SEVIR) data set (Veillette et al., 2020). It integrates data from the GOES-16 satellite and NEXRAD weather radars, capturing over 10,000 storm events. Each event comprises 4 hr of data at 5-min intervals, covering a 384 × 384 km region with a spatial resolution of 1 km. This data set has been widely utilized for developing and evaluating ML models for precipitation nowcasting (Gavahi et al., 2023; Yang & Yuan, 2023; Zheng et al., 2024).

In our study, the vertically integrated liquid (VIL) from SEVIR is used as the sole representation of precipitation. To enhance training efficiency, the temporal duration of VIL images is reduced to 2 hrs per event by removing the first and last hour of observations. The image frames are sampled every 10 min, while preserving the original spatial scope and resolution. This strategy preserves the effective sample size of the original training data set while ensuring comprehensive coverage of the entire lifecycle of convective storm events. A similar approach has been adopted in previous studies, including Gao et al. (2022) and Veillette et al. (2020). VIL pixel values are stored as integers ranging from 0 to 254. Only the VIL data set is utilized as both input and out for training and testing the ML models. The training data set comprises 5,088 storm events spanning 2016–2018, while the testing data set includes 1,872 events from 2019 to 2020. The task is formulated as predicting the next 1 hr of precipitation evolution, conditioned on observations from the preceding 1 hr. This adjustment shortens both the input sequence length and the forecast lead time; however, it does not affect the effective sample size of the training data set.

2.2. Data Set for Short-Term Precipitation Forecast

For the short-term precipitation forecasting task, we provide a brief overview of the data set used, with comprehensive details available in Cao et al. (2025). The ML model developed for this task leverages a multi-source data set integrating atmospheric variables across multiple vertical levels. The data sets used for the training and testing include satellite observations, ERA5 reanalysis data, Global Precipitation Measurement (GPM) rainfall data, and the China Meteorological Administration (CMA) Land Data Assimilation System ground data set. The model aims to forecast the instantaneous rainfall rate for the next 12 hr. As inputs, it utilizes the preceding 3 hr of GPM rainfall data, alongside the aforementioned data sets. Model training is conducted using data from 2018 to 2022, with 20% of the samples reserved for validation. The model's performance is evaluated on data from the year 2023. It should be noted that, for real-time applications, the ERA5 data set can be

replaced with high-resolution products generated by regional NWP systems operated by CMA, which are available in real time.

3. Methodology

3.1. Conventional Loss Function

The underestimation of extreme precipitation by ML models can, in part, be attributed to the design of loss function during training. These models typically employ loss functions that measure the aggregated differences between the predicted and observed precipitation values across geographical grid points, such as the MSE-based loss, as shown below:

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_i'\right)^2, \tag{1}$$

where N is the total number of grid points, y_i and y_i' represent the i th grid point from observation and model's prediction, respectively.

The forecasting model is trained by minimizing the loss function (Equation 1). Theoretically, minimizing this grid-point-wise loss function leads the model to approximate the expected value of the predicted variables, assuming a Gaussian distribution of prediction errors (Mlotshwa et al., 2022). While this approach is effective in capturing the large-scale precipitation structure, it is susceptible to the "double penalty" problem, and thus tends to smooth the smaller-scale precipitation structures and underestimate the heavy rainfall in observations (Mathieu et al., 2016). As a result, the ML models using the MSE-based loss function generally underestimate the frequency and intensity of heavy rainfall while overestimate those of moderate to light rainfall (see Figure 1a).

3.2. Probability-Matching Based Loss Function

Inspired by the concept of the PM-EM, we incorporate the PM constraint into MSE loss function, called PM loss function. The PM constraint explicitly considers the difference in the frequency distribution of the intensity of gridded precipitation between model forecast and ground truth. The fitting of these two frequency distributions is integrated into the conventional MSE loss function as an additional constraint. The implementation of this PM-based loss function is straightforward and involves the following steps (Figure 1b):

1. *Sorting*. Within each pair of forecast and observed precipitation fields ($H \times W$) in each sample of the training data set, the grid points of the prediction or observation fields are sorted in descending or ascending order based on precipitation intensity. This step establishes a one-to-one correspondence between forecast and observed grid points ranked by intensity, independent of their geographic location.
2. *Matching*. For the sorted sequences of precipitation forecast and observation in descending or ascending order, calculate the MSE of precipitation values between the pair-wise pixels:

$$\mathcal{L}_{\mathrm{PM}} = \frac{1}{N}\sum_{i=1}^{N}\left(z_i - z_i'\right)^2, \tag{2}$$

where z_i and z_i' are defined as the i th largest element in the sorted forecast and observation fields ($Z = \mathrm{sort}(Y)$ and $Z' = \mathrm{sort}(Y')$).

The PM component (i.e., $\mathcal{L}_{\mathrm{PM}}$) is then combined with the conventional MSE (i.e., $\mathcal{L}_{\mathrm{MSE}}$) as an additional constraint. The new PM loss function is then calculated as shown below:

$$\mathcal{L} = \mathcal{L}_{\mathrm{MSE}} + \omega \times \mathcal{L}_{\mathrm{PM}}, \tag{3}$$

where ω is the tunable scalar weighting parameter to scale the importance of the PM constraint, and will be tested and discussed in Section 3.4.

3.3. Advantages of PM Loss Function Compared to Existing Methods

Various strategies have been proposed to address the underestimation of extreme precipitation in forecasts when using the MSE loss function. These approaches generally involve reweighting precipitation intensities (e.g., Weighted MSE, WMSE; Balanced MSE, BMSE), applying scaling, or incorporating localization properties of precipitation (e.g., Ascenso et al., 2024) during training. Among these, two of the most effective and widely adopted methods are WMSE and BMSE (Nooten et al., 2023; Wang et al., 2022), which assign intensity-dependent weights within the loss function (see their specific schemes in Section S3 of Supporting Information S1). Despite their effectiveness, both WMSE and BMSE suffer from two major limitations. First, their formulations remain fundamentally anchored to the positionally matched MSE, which is prone to the "double penalty" problem and can compromise balanced performance across multiple verification metrics. Second, both require empirical tuning of multiple weighting parameters, which complicates the generalization across different models and forecasting scenarios.

To mitigate these challenges, the PM-based loss function integrates a traditional positionally matched MSE component with a position-agnostic constraint, governed by a single tunable weight (ω). The former enhances spatial accuracy in precipitation forecasts, while the latter promotes consistency between the frequency distributions of predicted and observed precipitation intensities. This dual-objective design enables more balanced optimization of overall precipitation forecast performance across all intensity categories and diverse verification metrics.

3.4. Practical Implementation of PM-Based Loss Function

In this study, we evaluate the feasibility and effectiveness of the PM loss function for both nowcasting and short-term forecasting of precipitation using various ML model architectures. Specifically, we adopt two representative models for precipitation nowcasting: Simpler Yet Better Video Prediction (SimVP) and Convolutional Long Short-Term Memory (ConvLSTM). Both SimVP and ConvLSTM are state-of-the-art models that leverage classical CNN and RNN architectures, respectively (Gao et al., 2022; Shi et al., 2015). Details regarding the training procedures of these two models on the SEVIR data set are provided in Gao et al. (2022). For short-term precipitation forecasting, an ML model was developed at the Shanghai Central Meteorological Observatory. This model employs a Vision Transformer architecture, which enables the efficient integration of input data sets with varying spatiotemporal resolutions (see Section 2.2 and Cao et al., 2025 for more details).

The adoption of the PM-based loss function offers two primary advantages. First, the PM loss is fully differentiable, enabling direct optimization through gradient backpropagation during training (see Section S2 in Supporting Information S1 for detailed explanation of differentiability). Second, the PM constraint can be seamlessly integrated into the loss function module without altering the core architecture of the model. This integration requires only minimal code modification (see the pseudocode in Table S1 of Supporting Information S1). As described in Equation 3, the PM-based loss function includes a tunable weighting parameter ω, which controls the relative importance of the PM constraint. In our experiments, ω was optimized based on the performance of CSI and Bias scores for moderate level of precipitation and above and ultimately set to 10 for both the nowcasting and short-term forecasting models (see the quantitative optimization in Section S5 of Supporting Information S1).

4. Results

In this section, we quantitatively evaluate the performance of models trained using the PM-based loss function in comparison to those trained with conventional MSE and its weighted variants, including WMSE and BMSE, under both nowcasting and short-range precipitation forecasting scenarios. As the evaluation of short-range precipitation forecasting yields qualitatively similar results to those observed in the nowcasting setting, we focus on presenting detailed nowcasting results, along with the frequency distribution for short-range forecasting in the main text. The remaining results for short-range forecasting are provided in the Section S7 of Supporting Information S1. BMSE's performance is generally comparable to the WMSE's, and we have included its results in the SI for brevity.

4.1. Forecast Skill Evaluation

Given the similarity in performance between the SimVP and ConvLSTM models, only the results of ConvLSTM are presented in the main text, while those of SimVP are provided in the Section S6 of Supporting Information S1. We begin by comparing error-based metrics across different models, that is, RMSE (see Figure S3 in Supporting Information S1 for details). As expected, the model trained solely with MSE loss exhibits the lowest forecast error, which can be attributed to the direct alignment between the optimization objective and the evaluation metric. In contrast, the PM-based loss yields slightly higher errors, as it incorporates an additional PM constraint alongside the MSE loss. A higher weight on the PM constraint leads to a higher RMSE (cf. PM($\omega = 1$) and PM ($\omega = 10$)). Among all models, the one trained with WMSE-loss demonstrates the poorest performance in terms of RMSE. This discrepancy may be attributed to the model's assignment of higher weights to moderate-to-heavy precipitation events with substantial variability during the training phase. These findings highlight an inherent trade-off between enhancing forecast accuracy (or resolution) for extreme precipitation events and minimizing conventional error metrics—a manifestation of the well-known double-penalty problem.

In addition to error-based metrics, we employ threshold-based binary metrics to evaluate nowcasting skill across different precipitation intensity categories. The pixel value thresholds of 74, 100, and 133 used for the SEVIR data set approximately correspond to the total liquid water in cloud of 0.78, 1.5, and 3.5 kg/m^2, respectively, as derived from the empirical equation in Veillette et al. (2020). These thresholds are used to roughly categorize precipitation intensity for nowcasting into three levels—light, moderate, and heavy. The selection of these precipitation thresholds follows the conventions adopted by the majority of ML models evaluated on the SEVIR data set (e.g., Gao et al., 2022; Yang & Yuan, 2023). The evaluation metrics include the CSI, POD, FAR, and BIAS (see Section S4 in Supporting Information S1 for more details of their algorithms), which are widely used to assess a model's ability to capture precipitation at specific thresholds.

The performance of WMSE-, MSE-, and PM-based loss functions is generally consistent across light, moderate, and heavy rainfall (Figure 2). For light and moderate precipitation, all three models yield comparable CSI values. For heavy precipitation, the WMSE and PM ($\omega = 10$) achieve slightly higher CSI than PM ($\omega = 1$), followed by MSE loss. This improvement can be attributed to the explicit weighting of heavy rainfall in both the WMSE and PM loss functions, albeit through different strategies. In terms of POD, the WMSE-based model performs best, followed by the PM- and then MSE-based models; this trend is reversed for FAR, with WMSE and MSE exhibiting overall the worst and best performance, respectively. These results are consistent with the BIAS characteristics of each model: WMSE exhibits a pronounced positive bias due to its empirical weighting emphasizing higher intensity, while MSE shows a strong negative bias. The positive bias of WMSE possibly reduces the rate of missed detections, resulting in higher POD and CSI. However, this same bias also tends to inflate false alarms, thereby increasing FAR score of WMSE (see Equations 8–11 in Section S4 of Supporting Information S1). While the MSE-loss model performs best on the FAR metric, this is primarily due to its substantial negative bias.

Among the three, the PM-loss model demonstrates a relatively more balanced performance across all metrics. It achieves BIAS scores closest to unity across VIL pixel intensities, indicating an accurate representation of precipitation. This improvement stems from the PM constraint in training, effectively reducing systematic bias. With an appropriately chosen weight ($\omega = 10$), the PM-loss model attains POD and FAR scores that lie between those of the WMSE- and MSE-based models, while maintaining a relatively high CSI score. These results are similar for evaluation of short-term precipitation forecasts (see Figure S11 in Section S7 of Supporting Information S1).

4.2. Frequency Distribution and Power Spectrum of Precipitation

Beyond forecast skill, assessing the realism—or reliability—of precipitation forecasts is crucial. We evaluate this by analyzing the power spectrum and frequency distribution of predicted precipitation fields. Figure 3 shows the power spectral density at 1-hr lead time for the SimVP (Figure 3a) and ConvLSTM (Figure 3b) model trained with MSE, WMSE, and PM losses (with PM constraint weights $\omega = 1$, 10, and 100). All models underestimate the spectral power, particularly at scales below 500 km, due to the MSE component of the loss function that suppresses small-scale variability over time. In contrast, PM and WMSE losses better preserve spectral power, indicating improved retention of intensity features. Increasing the PM constraint weight to $\omega = 100$ substantially

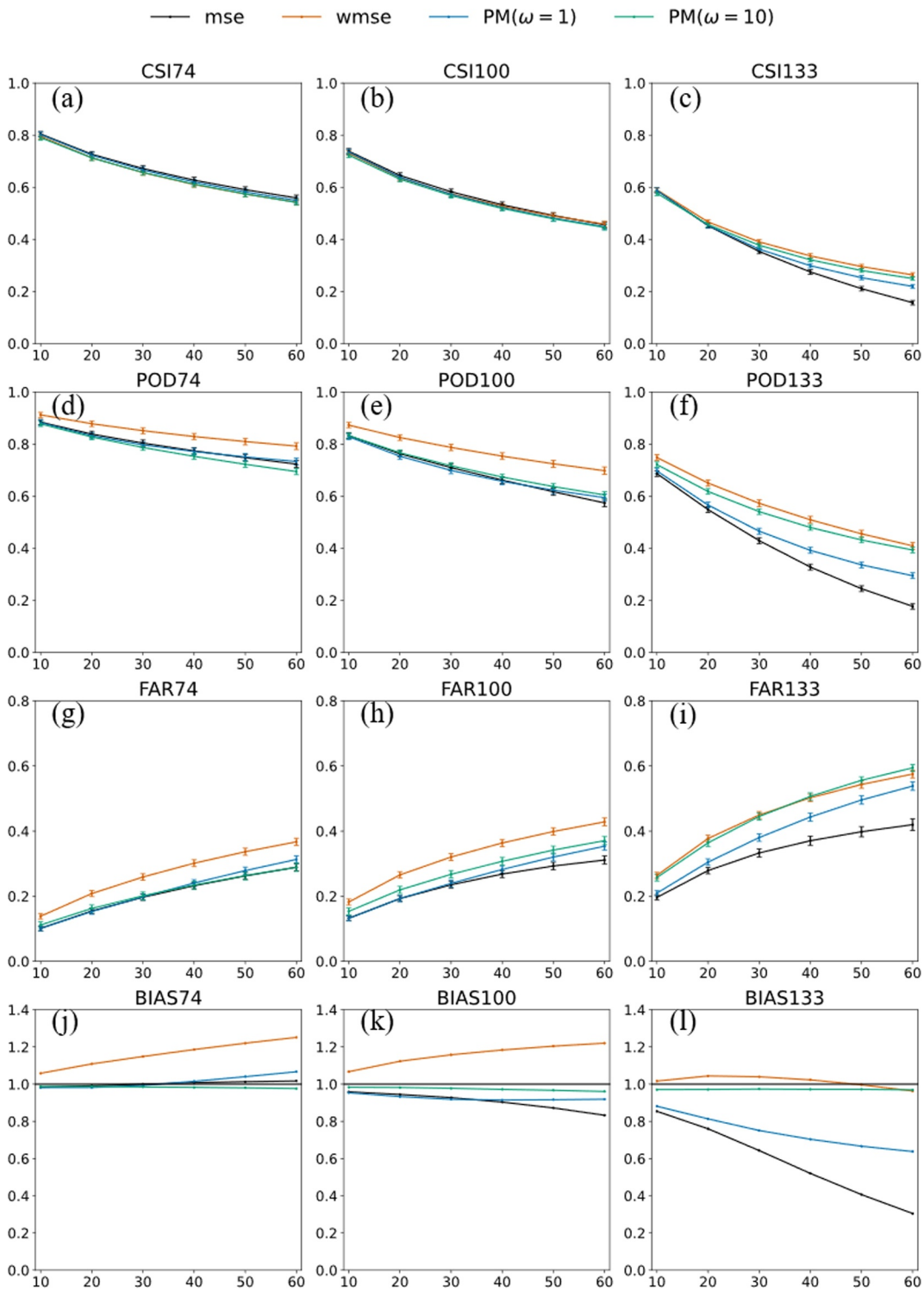

Figure 2. Quantitative assessment of MSE-, WMSE-, and PM-based ConvLSTM for 1-hr precipitation nowcasting with threshold (74, 100, and 133), depicted as functions of lead time. Panels (a–c), (d–f), (g–i), (j–l) are the CSI, POD, FAR, and BIAS scores with specific threshold, respectively. The significance test is conducted using a *t*-test with a 99% confidence level and the error bar represents confidence interval. The numerical values of 74,100,133 represent specific precipitation intensities within the 0–255 pixel value range and correspond to 0.78, 1.51, and 3.53 kg/m^2, respectively.

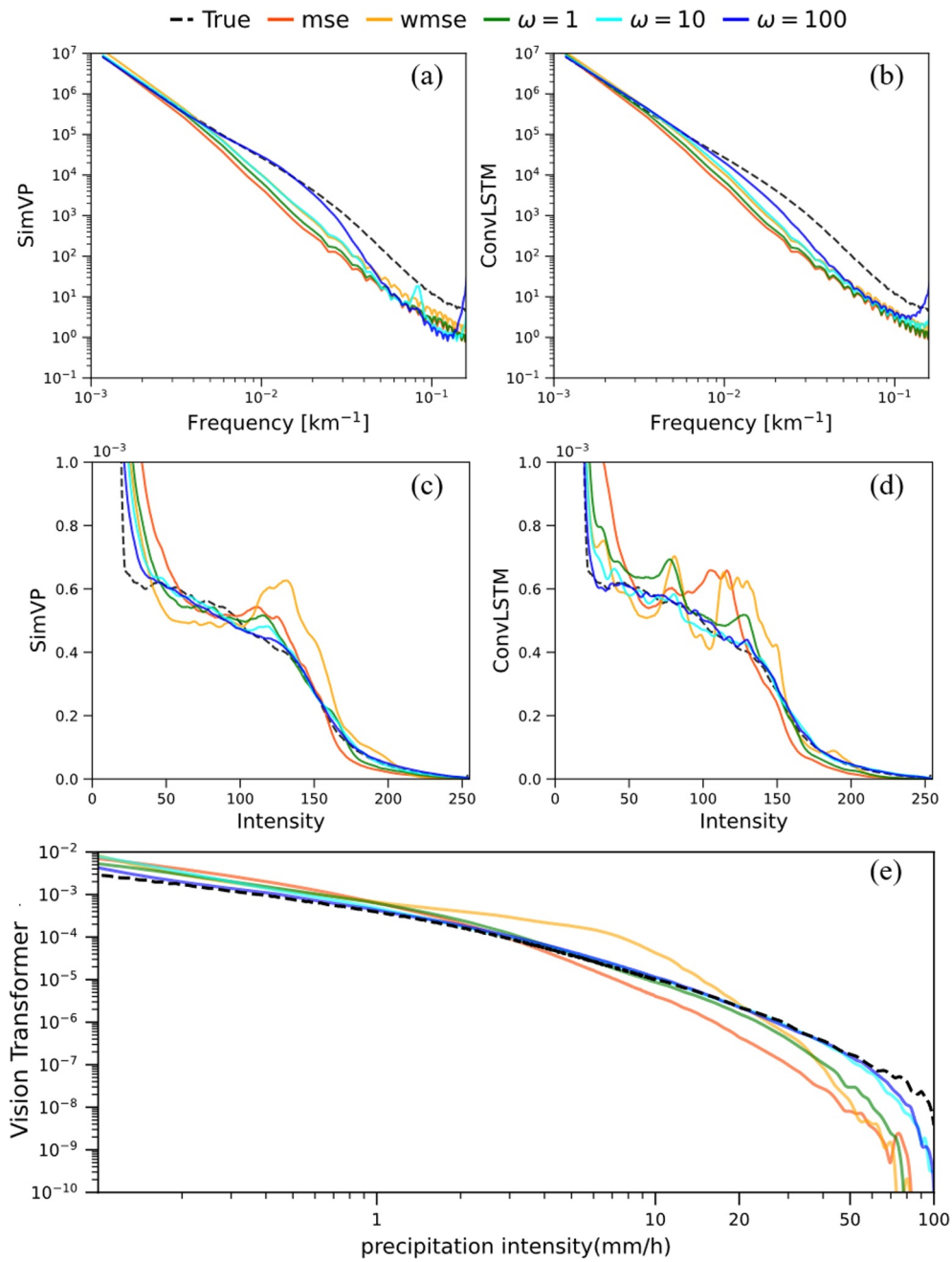

Figure 3. The power spectral density (a, b) and frequency distribution (c, d) of MSE-WMSE-,and PM-based for ML models (SimVP and ConvLSTM) in precipitation nowcasting at 1-hr lead time. The frequency distribution of short-term precipitation forecast is also shown in panel (e). Black dashed line denotes observational data, red and orange solid lines indicate MSE- and WMSE-based models, respectively. The green, cyan, and blue solid lines correspond to PM-based models with weighting factors of 1, 10, and 100, respectively.

improves the spectral alignment with the reference. However, this enhancement may come at the cost of reduced overall forecast accuracy (see Figures S1 and S2 in Section S5 of Supporting Information S1).

Similarly, the analysis of the frequency distribution of precipitation demonstrates that the PM-based loss function with weights $\omega = 10$ and 100 produce very similar results, both of which substantially outperform the WMSE and MSE losses in matching the reference frequency distribution for SimVP (Figure 3c) and ConvLSTM (Figure 3d) of nowcasting and the short-term forecast (Figure 3e). This is particularly notable for moderate rainfall levels

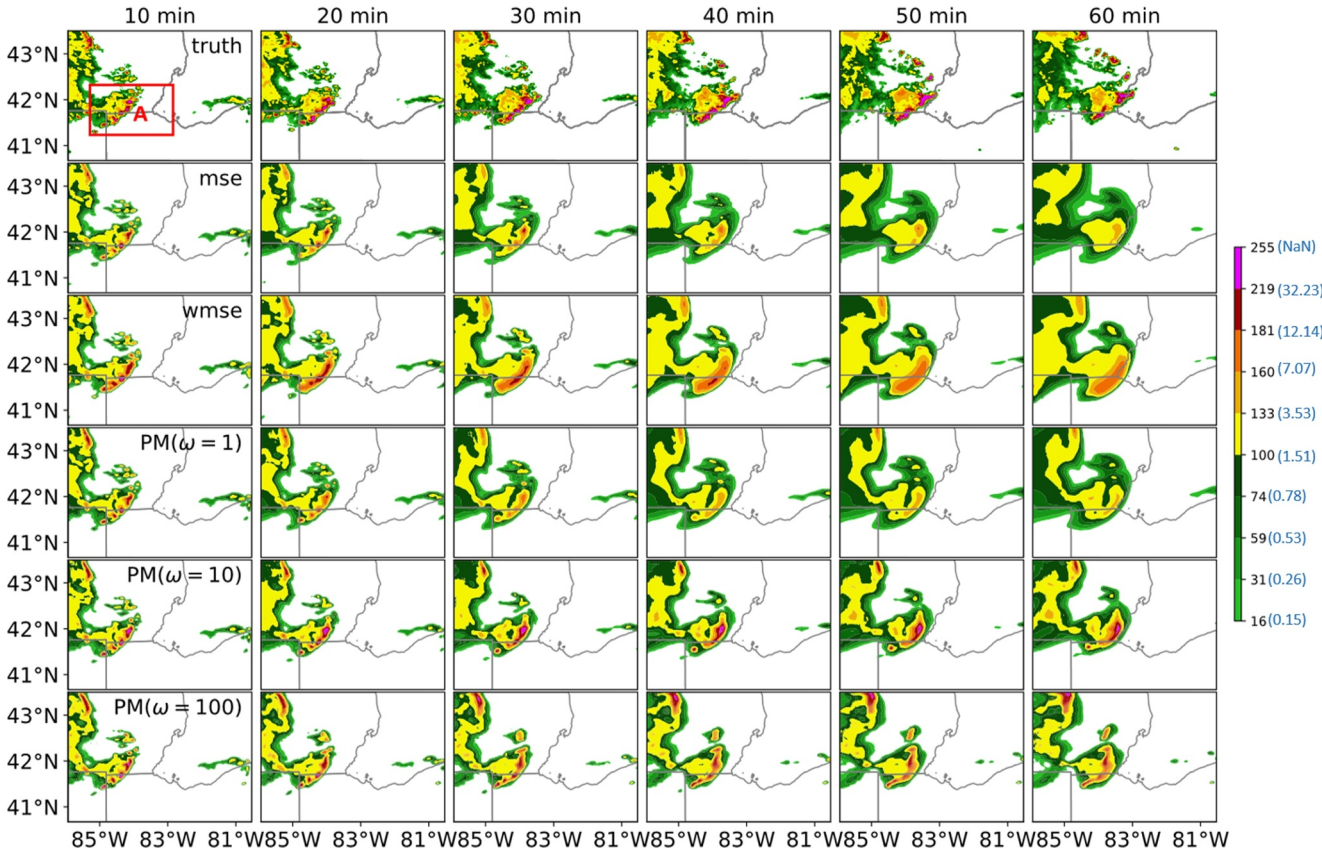

Figure 4. Comparative analysis of MSE-, WMSE-, PM-based ConvLSTM in 1-hr precipitation nowcasting: a case study (event ID:S850653). Rows sequentially display: (1) observational data; (2) MSE-based predictions; (3) WMSE-based predictions; (4–6) PM-based predictions with weighting factors of 1, 10, 100, respectively. The black and blue annotations of legend correspond to pixel and physicial units (kg/m^2), respectively.

(VIL pixel value >50 for nowcasting and precipitation rate >3 mm/hr for short-term forecast). In contrast, the MSE loss significantly overestimates the frequency of light rainfall while underestimating heavier events, Meanwhile, the WMSE loss exaggerates heavy rainfall frequencies due to its design of additional weight.

4.3. Case Study

To complement the large-sample evaluation, we analyze a representative nowcasting case (Figure 4) from a thunderstorm event on 18 August 2019 (16:23), featuring a combination of stratiform cloud to the west and structured cell storms to the east. Due to similar model performance within the first 30 min, we focus on lead times beyond 30 min. The MSE-loss-based ConvLSTM captures the storm's shape, motion, and key precipitation centers but severely underestimates intensity and overestimates light rainfall extent. This is due to the inherent smoothing effect of the MSE loss.

WMSE- and PM-loss-based ConvLSTM models mitigate these issues with differing trade-offs. WMSE-loss improves heavy rainfall prediction but overestimates moderate-to-heavy precipitation areas, leading to higher FAR due to its elevated bias as indicated in Figure 2. In contrast, PM-loss maintains intensity persistence and accurately captures the spatial distribution of heavy rainfall, benefiting from balanced bias across intensities. Adjusting the PM constraint weight from 1 to 10 improves predictions, particularly for heavy rainfall center "A." At weight 100, the forecast captures finer structures but suffers spatial displacement of heavy precipitation due to overemphasis on frequency alignment at the cost of positional accuracy. We also analyzed two additional cases involving scattered supercell storms with differing intensities and spatial densities (see Figures S5–S9 in Section S6 of Supporting Information S1) and a typical short-term forecast case of storm rainfall (see Figure S12 in Section S7 of Supporting Information S1), and all cases yield qualitatively consistent conclusions.

5. Conclusions

Despite notable advancements, precipitation nowcasting and short-term forecasting using ML frameworks still face significant challenges in accurately predicting heavy rainfall events. A critical limitation stems from the widespread use of MSE-based loss functions in training ML models. These loss functions tend to underestimate the intensity of heavy rainfall due to the so-called "double penalty" effect. To address this issue, we propose a novel probability matching (PM)-based loss function. This new approach integrates a constraint into the conventional MSE framework that enforces consistency between the frequency distributions of predicted and observed precipitation. Specifically, the PM constraint is implemented by calculating the aggregated error between the frequency distributions of the predicted and observed precipitation fields. To rigorously assess the effectiveness of the PM-based loss function, we evaluate its forecasting skill relative to both WMSE- and standard MSE-based models. The WMSE is a reweighted MSE in which heavy precipitation samples are assigned greater weight during training, thereby enhancing the model's skill in predicting heavy rainfall. Experiments are conducted across multiple ML architectures and for different forecasting horizons, including nowcasting (1 hr) and short-term forecasting (12 hr).

A comprehensive evaluation of forecasting skill reveals that the relative performance of models trained with PM-, WMSE-, and MSE-based loss functions is generally similar for both nowcasting and short-term forecasting tasks. The PM-based loss function results in a slightly higher RMSE compared to the MSE-based model, primarily due to its reduced emphasis on the MSE component within the composite loss function. The WMSE-based model produces the highest RMSE, as it assigns greater weight to high-intensity precipitation, thereby amplifying errors in those regions.

To complement this evaluation, we further assess performance using threshold-based binary metrics. The results highlight a key strength of the PM-based loss function: it yields the lowest precipitation forecast bias across all rainfall intensities, from light to heavy. As a result, the PM-based model achieves a significantly lower FAR compared to the WMSE-based model. While the WMSE-based model demonstrates superior performance in terms of the POD, this is largely driven by its pronounced positive bias. Conversely, the MSE-based model exhibits a strong negative bias across all rainfall categories, resulting in the best FAR but the worst POD. In terms of CSI, all models perform comparably. The PM-based and WMSE-based models show similar CSI score for heavy rainfall which is slightly better than the MSE loss. Overall, the PM-based loss function achieves relatively more balanced and consistent performance across all evaluation metrics compared to other loss functions.

To further evaluate forecast reliability, we examined both the power spectral density and the frequency distribution of 1-hr nowcasting and 12-hr short-term forecasting of precipitation. Relative to the MSE-based model, the PM-based model retains markedly more spectral power at spatial scales <500 km, indicating superior preservation of small-scale precipitation variability as forecasts evolve. The frequency analysis paints a consistent picture. When the PM term is given a weight of 10, the resulting frequency distribution aligns most closely with the observed climatology: the model reduces the over-prediction of light rainfall, appropriately boosts moderate and heavy events, and avoids the excessive inflation of extreme-rainfall frequencies observed in the WMSE-based model. Case-study diagnostics of some typical convective storm events further underscore these advantages.

The above findings demonstrate that the PM-based loss function delivers consistent gains in both predictive skill and reliability across multiple ML architectures and forecasting horizons. Moreover, the PM constraint is conceptually transferable to other forecasting contexts—particularly those involving spatially intermittent fields such as hail or fog—where matching the observed occurrence distribution is critical. Nevertheless, when the precipitation field exhibits multiple spatially separated centers, the use of a global PM approach may introduce artifacts or distortions. Developing a localized PM scheme could therefore be a promising direction for future investigation.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The Storm Event ImagRy (SEVIR) data set for nowcasting is available on the Amazon Web Services S3 (Veillette et al., 2020). The code for SimVP and ConvLSTM in precipitation nowcasting can be found in Gao et al. (2022). For the short-term precipitation forecasting, the radar data set is available on the Zenodo (Chen, 2024). The ERA5 data are available at Copernicus Climate Data Store (Hersbach et al., 2018). The satellite data can be accessed at the following steps: ftp://w4c@ala.boku.ac.at/. The access phrase required is "Weather4cast23". For further details on data access, please visit the official webset: https://weather4cast.net/neurips2024/.

Acknowledgments
This research was jointly supported by Open project of Key Laboratory of High Impact Weather (special), China Meteorological Administration (2025-K-03), the Shanghai Municipal Science and Technology Commission Scientific Research Program, the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (2025-GZL-RGZN-BTBX-02034), Guidance Project for Industrial Technology Development and Application Plan of Fujian Province (NO.2024Y0075), the Joint Research Project for Meteorological Capacity Improvement (24NLTSZ004), and the National Natural Science Foundation of China (Grant 42105001).

References

Arcomano, T., Szunyogh, I., Pathak, J., Wikner, A., Hunt, B. R., & Ott, E. (2020). A machine learning-based global atmospheric forecast model. *Geophysical Research Letters*, *47*(9), e2020GL087776. https://doi.org/10.1029/2020GL087776

Ascenso, G., Ficchì, A., Giuliani, M., Scoccimarro, E., & Castelletti, A. (2024). Downscaling, bias correction, and spatial adjustment of extreme tropical cyclone rainfall in ERA5 using deep learning. *Weather and Climate Extremes*, *46*, 100724. https://doi.org/10.1016/j.wace.2024.100724

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Bowler, N. E., Pierce, C. E., & Seed, A. W. (2006). STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quarterly Journal of the Royal Meteorological Society*, *132*(620), 2127–2155. https://doi.org/10.1256/qj.04.100

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2021). Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200086. https://doi.org/10.1098/rsta.2020.0086

Bremnes, J. B. (2004). Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Monthly Weather Review*, *132*(1), 338–347. https://doi.org/10.1175/1520-0493(2004)132<0338:pfopit>2.0.co;2

Cao, Y., Chen, L., Wu, J., & Feng, J. (2025). Enhancing nowcasting with multi-resolution inputs using deep learning: Exploring model decision mechanisms. *Geophysical Research Letters*, *52*(4), e2024GL113699. https://doi.org/10.1029/2024GL113699

Chen, L. (2024). Radar reflectivity [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.12749010

Ebert, E. E. (2001). Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review*, *129*(10), 2461–2480. https://doi.org/10.1175/1520-0493(2001)129<2461:aoapms>2.0.co;2

Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C., et al. (2022). Deep learning for twelve hour precipitation forecasts. *Nature Communications*, *13*(1), 5145. https://doi.org/10.1038/s41467-022-32483-x

Feng, J., Zhang, J., Toth, Z., Peña, M., & Ravela, S. (2020). A new measure of ensemble central tendency. *Weather and Forecasting*, *35*(3), 879–889. https://doi.org/10.1175/WAF-D-19-0213.1

Gao, Z., Shi, X., Han, B., Wang, H., Jin, X., Maddix, D., et al. (2023). PreDiff: Precipitation nowcasting with latent diffusion models. *arXiv*. https://doi.org/10.48550/arXiv.2307.10422

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

Gao, Z., Tan, C., Wu, L., & Li, S. Z. (2022). SimVP: Simpler yet better video prediction. In *Presented at the proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3170–3180).

Gavahi, K., Foroumandi, E., & Moradkhani, H. (2023). A deep learning-based framework for multi-source precipitation fusion. *Remote Sensing of Environment*, *295*, 113723. https://doi.org/10.1016/j.rse.2023.113723

Germann, U., Galli, G., Boscacci, M., & Bolliger, M. (2006). Radar precipitation measurement in a mountainous region. *Quarterly Journal of the Royal Meteorological Society*, *132*(618), 1669–1692. https://doi.org/10.1256/qj.05.190

Groenemeijer, P., Púčik, T., Holzer, A. M., Antonescu, B., Riemann-Campe, K., Schultz, D. M., et al. (2017). Severe convective storms in Europe: Ten years of research and education at the European severe storms laboratory. *Bulletin of the American Meteorological Society*, *98*(12), 2641–2651. https://doi.org/10.1175/BAMS-D-16-0067.1

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200092. https://doi.org/10.1098/rsta.2020.0092

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., et al. (2018). ERA5 hourly data on single levels from 1940 to present [Dataset]. *Copernicus Climate Change Service (C3s) Climate Data Store (Cds)*, *10*(10.24381). https://doi.org/10.24381/cds.adbb2d47

Hu, Y., Yin, F., & Zhang, W. (2021). Deep learning-based precipitation bias correction approach for Yin–He global spectral model. *Meteorological Applications*, *28*(5), e2032. https://doi.org/10.1002/met.2032

Imhoff, R. O., Brauer, C. C., Overeem, A., Weerts, A. H., & Uijlenhoet, R. (2020). Spatial and temporal evaluation of Radar rainfall nowcasting techniques on 1,533 events. *Water Resources Research*, *56*(8), e2019WR026723. https://doi.org/10.1029/2019WR026723

Kim, H., Ham, Y. G., Joo, Y. S., & Son, S. W. (2021). Deep learning for bias correction of MJO prediction. *Nature Communications*, *12*(1), 3087. https://doi.org/10.1038/s41467-021-23406-3

Leutbecher, M., Lock, S.-J., Ollinaho, P., Lang, S. T. K., Balsamo, G., Bechtold, P., et al. (2017). Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, *143*(707), 2315–2339. https://doi.org/10.1002/qj.3094

Lin, C., Vasić, S., Kilambi, A., Turner, B., & Zawadzki, I. (2005). Precipitation forecast skill of numerical weather prediction models and radar nowcasts. *Geophysical Research Letters*, *32*(14), 2005GL023451. https://doi.org/10.1029/2005GL023451

Mathieu, M., Couprie, C., & LeCun, Y. (2016). Deep multi-scale video prediction beyond mean square error. *arXiv*. https://doi.org/10.48550/arXiv.1511.05440

McGovern, A., Chase, R. J., Flora, M., Gagne, D. J., Lagerquist, R., Potvin, C. K., et al. (2023). A review of machine learning for convective weather. *Artificial Intelligence for the Earth Systems*, *2*(3), e220077. https://doi.org/10.1175/AIES-D-22-0077.1

Mlotshwa, T., van Deventer, H., & Bosman, A. S. (2022). Cauchy loss function: Robustness under Gaussian and Cauchy noise. In A. Pillay, E. Jembere, & A. Gerber (Eds.), *Artificial intelligence research* (pp. 123–138). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-22321-1_9

Pierce, C., Seed, A., Ballard, S., Simonin, D., & Li, Z. (2012). Nowcasting. In J. Bech & J. L. Chau (Eds.), *Doppler Radar observations-weather radar, wind profiler, ionospheric radar, and other advanced applications*. IntechOpen. https://doi.org/10.5772/39054

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., & Foresti, L. (2019). Pysteps: An open-source python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, *12*(10), 4185–4219. https://doi.org/10.5194/gmd-12-4185-2019

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677. https://doi.org/10.1038/s41586-021-03854-z

Sattari, A., Foroumandi, E., Gavahi, K., & Moradkhani, H. (2025). A probabilistic machine learning framework for daily extreme events forecasting. *Expert Systems with Applications*, *265*, 126004. https://doi.org/10.1016/j.eswa.2024.126004

Seed, A. W., Pierce, C. E., & Norman, K. (2013). Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme: Formulation of a scale decomposition nowcast scheme. *Water Resources Research*, *49*(10), 6624–6641. https://doi.org/10.1002/wrcr.20536

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., & WOO, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* (Vol. 28). Curran Associates, Inc.

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W., et al. (2017). Deep learning for precipitation nowcasting: A benchmark and A new model. In *Advances in Neural Information Processing Systems*. (Vol. 30(8), pp. 743–750). https://doi.org/10.3969/j.issn.1003-0034.2017.08.013

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., et al. (2020). MetNet: A neural weather model for precipitation forecasting. *arXiv*. https://doi.org/10.48550/arXiv.2003.12140

Toth, E., Brath, A., & Montanari, A. (2000). Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology*, *239*(1–4), 132–147. https://doi.org/10.1016/S0022-1694(00)00344-9

van Nooten, C. C., Schreurs, K., Wijnands, J. S., Leijnse, H., Schmeits, M., Whan, K., & Shapovalova, Y. (2023). Improving precipitation nowcasting for high-intensity events using deep generative models with balanced loss and temperature data: A case study in the Netherlands. *Artificial Intelligence for the Earth Systems*, *2*(4), e230017. https://doi.org/10.1175/AIES-D-23-0017.1

Veillette, M., Samsi, S., & Mattioli, C. (2020). SEVIR: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In *Advances in neural information processing systems* (Vol. 33, pp. 22009–22019).

Wang, C., Wang, P., Wang, P., Xue, B., & Wang, D. (2022). A spatiotemporal attention model for severe precipitation estimation. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2021.3084293

Wang, R., Fung, J. C. H., & Lau, A. K. H. (2023). Physical-dynamic-driven AI-synthetic precipitation nowcasting using task-segmented generative model. *Geophysical Research Letters*, *50*(21), e2023GL106084. https://doi.org/10.1029/2023gl106084

Xu, X., Liu, Y., Chao, H., Luo, Y., Chu, H., Chen, L., et al. (2019). Towards a precipitation bias corrector against noise and maldistribution. *arXiv*. https://doi.org/10.48550/arXiv.1910.07633

Yang, S., & Yuan, H. (2023). A customized multi-scale deep learning framework for storm nowcasting. *Geophysical Research Letters*, *50*(13), e2023GL103979. https://doi.org/10.1029/2023GL103979

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., & Wang, J. (2023). Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, *619*(7970), 526–532. https://doi.org/10.1038/s41586-023-06184-4

Zheng, K., He, L., Ruan, H., Yang, S., Zhang, J., Luo, C., et al. (2024). A cross-modal spatiotemporal joint predictive network for rainfall nowcasting. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–23. https://doi.org/10.1109/TGRS.2024.3452767