Improving the fine structure of intense rainfall forecast by a designed adversarial generation network

Zuliang Fang[1], Qi Zhong[1], Haoming Chen[2], Xiuming Wang[1], Zhicha Zhang[3], and Hongli Liang[1]

[1]China Meteorological Administration Training Center,Beijing 100081, China
[2]Chinese Academy of Meteorological Sciences, Beijing 10081, China
[3]Zhejiang Meteorological Observatory,Hangzhou 310017, China

Correspondence: Qi Zhong (zhongq@cma.gov.cn)

Abstract. Accurate short-term precipitation forecasting is critical for socio-economic activities. However, due to inherent deficiencies of numerical weather prediction (NWP) models, the accuracy of precipitation forecasts ~~remains~~ is significantly inadequate. In recent years, Deep learning (DL) has been employed to enhance precipitation forecasts, yet these forecasts frequently appear blurry and fail to meet the precision required for operational applications. In this paper, we propose a Generative Adversarial Fusion Network (GFRNet) ~~designed~~ to provide quantitative forecasts of 3-hour accumulated precipitation over the next 24 hours in North China, based on the outputs ~~of~~ from multiple ~~numerical weather~~ NWP models. Evaluation results indicate that GFRNet outperforms ~~numerical models~~ NWP models across all precipitation intensities. Specifically, GFRNet's threat scores (TS) improved by 4%, 28%, 35%, and 19% at thresholds of 0.1 mm, 10 mm, 20 mm, and 40 mm, respectively, compared to the highest spatial resolution regional numerical model of the China Meteorological Administration (CMA-3KM). Additionally, GFRNet's Fraction Skill Scores (FSS) at thresholds of 10 mm, 20 mm, and 40 mm show improvements of 13%, 18%, and 15% respectively. ~~over those of CMA-3KM.~~ These enhancements are consistent across most spatial regions and forecast lead times. Furthermore, ~~GFRNet outperforms all models in terms of Root Mean Square Error (RMSE) and Multi-Scale Structural Similarity Index (MS-SSIM).~~ GFRNet archives the best performance in Multi-Scale Structural Similarity Index (MS-SSIM) and ranks second in Root Mean Square Error (RMSE), significantly outperforming CMA-3KM. ~~Compared to the DL-based precipitation model FRNet, which lacks a generative strategy and tends to produce blurry forecasts with over-prediction~~ Compared to the DL-based model FRNet, which lacks a generative mechanism and often yields blurry predictions with overestimation, GFRNet better captures the fine structure and ~~evolutionary patterns~~ temporal evolution of precipitation, demonstrating significant operational value.

1 Introduction

Numerical Weather Prediction (NWP) serves as a fundamental tool in routine precipitation forecasting. However, its accuracy is ~~hindered by several~~ contrained by various factors, including initial condition errors, limited ~~model~~ spatial resolution, incomplete physical parameterizations, and approximate boundary conditions, all of which contribute to persistent forecast ~~errors~~ uncertaines (Sun et al., 2014; Boeing, 2016). As a result, it is challenging for any single numerical model to accurately capture the location, intensity, and structural evolution of precipitation.

1

25   ~~Deep learning, a key technology in artificial intelligence, has found numerous applications in NWP post-processing, large-scale data processing, super-resolution downscaling, and spatio-temporal forecasting~~In recent years, deep learning (DL), a core technique in artificial intelligence, has been increasingly applied in meteorology for NWP post-processing, large-scale data assimilation, super-resolution downscaling, and spatiotemporal prediction (Yang et al., 2022). ~~Significant research has also been conducted in the field of precipitation forecasting.~~In the domain of precipitation forecasting, DL has achieved signif-

30 icant progress. For nowcasting ~~task within 0-6 hours~~(0 - 6 hours), purely data-driven ~~deep learning~~ DL methods based on ~~minute-level~~ radar and satellite data have demonstrated substantial superiority over numerical models and optical flow methods(Shi et al., 2015; Wang et al., 2018b; Sønderby et al., 2020; Espeholt et al., 2022). For short-term forecasting within the 6-24 hour, precipitation prediction primarily relies on post-processing of ~~numerical model~~NWP outputs. For example, Zhang et al. (2020) ~~developed a precipitation correction model for the 12-h accumulated precipitation in eastern China using LSTM~~

35 ~~algorithms based on control forecast data from the European Center for Medium-Range Weather Forecasts (ECMWF) ensemble prediction system, the model showed~~developed an LSTM-based correction model for 12-hour accumulated precipitation over eastern China using ECMWF ensemble control forecasts, demonstrating superior performance both for light rain($< 5mm/12h$) and heavy rain($> 30mm/12h$) compared to frequency matching and SVM-based algorithms. Similarly, Chen et al. (2021) constructed an hourly precipitation correction model using a Convolutional Neural Network CNN based on mesoscale forecasts

40 from the East China Regional Numerical Center (CMA-SH9), which outperformed the probability matching method.

    Moreover, Zhou et al. (2022) utilized a 3D ~~convolutional network~~CNN to ~~model~~learn the nonlinear relationship between basic meteorological variables from the ECMWF's fifth-generation reanalysis dataset (ERA5) and corresponding 3-hour accumulated precipitation. ~~The trained model was then used to predict 3-hour accumulated precipitation by inputting basic variables from the ECMWF high-resolution model, showing significant improvement in the Threat Score (TS) at the $20mm/3h$~~

45 ~~threshold compared to the ECMWF forecast within 0-72 hours.~~ Their model, when applied to ECMWF high-resolution forecasts, significantly improved the Threat Score (TS) at the 20 mm/3h threshold for lead times up to 72 hours. In another study, Kim et al. (2022) used basic meteorological variables and precipitation from numerical model forecasts as input features for a ~~deep learning~~ DL model, achieving positive correction effects for light and moderate precipitation, though ~~it was less effective for~~the improvments diminished for precipitation ~~above~~exceeding 10 mm. Chen et al. (2023) employed ~~a weighted~~

50 ~~loss U-Net network~~a U-Net architecture with a weighted loss function to correct 6-hour accumulated precipitation predictions from the ECMWF, using 0.25° ERA5 precipitation data as a ~~target~~ground truth. This approach showed improvements across various precipitation intensities, from light rain ($\geq 0.1mm/6h$) to rainstorms ($\geq 20mm/6h$), in TS scores compared to the ECMWF forecast. Sun et al. (2023) developed a DABU-Net model combining data augmentation with deep learning to improve GFS wintertime precipitation forecasts over southeastern China. The model significantly enhanced Threat Scores

55 (TS) across multiple thresholds, with TS at the 20 mm/day threshold increasing by up to 100% at 72-hour lead time. Despite these advances, grid-based ~~deep learning~~ DL precipitation correction ~~methods typically achieve better results~~models generally perform better for light to moderate precipitation. ~~However, the improvement is less pronounced for heavy rainfall. Even when there is an improvement in TS scores for heavy rainfall, the predictions often appear overly smooth, lacking clear spatial precipitation structures, and the corresponding BIAS scores are often significantly greater than 1, which undermines the~~

60 ~~operational applicability of these algorithms.~~Improvements in TS for heavy rainfall are often accompanied by overly smoothed predictions, lacking well-defined spatial structures. Additionally, corresponding BIAS scores frequently exceed 1, indicating systematic overestimation and reducing the operational applicability of such methods.

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), as a typical deep generative model (DGM), have successfully transformed the intractable likelihood function into a neural network framework, enabling the model to optimize its

65 parameters to fit the likelihood function. ~~Through an adversarial process between the generator and the discriminator, GANs allow the generator to produce samples that closely resemble real data.~~By learning through competition between a generator and a discriminator, GANs enable the production of outputs that closely resemble the distribution of real data. GANs have been widely successful in image super-resolution tasks (Wang et al., 2018a) and have shown great promise in addressing challenges in short-term forecasting, such as excessive smoothing and the degradation of intensity over time (Ravuri et al., 2021;

70 Zhang et al., 2023). GANs have also demonstrated strong performance in statistical downscaling within the meteorological field (Leinonen et al., 2021; Price and Rasp, 2022; Singh et al., 2019). ~~There have been numerous successful applications of GANs in short-term precipitation forecasting and post-processing of NWP precipitation.~~Recent studies have explored the use of GANs in post-processing NWP-based precipitation forecasts. For example, Price and Rasp (2022) utilized a 4 km ~~spatial~~ resolution radar precipitation product ~~as a reference to construct a CGAN model for correcting and statistically downscaling the~~

75 ~~6-hour accumulated precipitation from the 32 km resolution ECMWF ensemble forecast.~~to train a conditional GAN (CGAN) model for correcting and downscaling 6-hour precipitation forecasts from the 32 km ECMWF ensemble. The ~~findings showed that the~~ CGAN model outperformed ~~a standard~~ CNN baselines and achieved comparable to high-resolution regional ensemble forecasts, ~~particularly~~especially for heavy precipitation events ($\geq$ 30mm/6h). Similarly, Harris et al. (2022) aimed to generate high-resolution ensemble precipitation forecasts by post-processing ECMWF forecasts at 10 km resolution using GAN and

80 VAE-GAN methods, targeting 1-hour accumulated precipitation products at 1 km resolution. Compared to traditional methods, the GAN approach showed significant advantages in preserving precipitation structure and predicting heavy precipitation ($\geq$ 5mm/1h). However, ~~these studies primarily use GANs to generate~~most existing applications of GANs focus on probabilistic ensemble forecasts rather than deterministic quantitative precipitation forecasts, and ~~they often do not focus on~~few studies directly address severe storm precipitation~~, which is of greater concern due to its hazardous nature.~~an area of critical operational

85 importance due to the associated risks.

Short-term heavy precipitation is typically characterized by sudden onset, short duration, small spatial scale, and high localization. These features demand precipitation forecasts with finer temporal and spatial resolutions to meet operational needs. In this study, we employ a GAN-based model, GFRNet, to generate deterministic forecasts of 3-hour accumulated precipitation over the next 24 hours in North China, using multiple NWP model outputs as input and targeting a resolution of 5 km.

90 Compared with previous research, this study introduces the following key advancements:

~~Due to the characteristics of precipitation, especially short-term heavy precipitation, which include sudden onset, short duration, small scale, and strong locality, a more refined temporal scale for deterministic quantitative precipitation forecasts holds greater operational significance. In this study, we employ GFRNet to predict 3-hour accumulated precipitation over the next 24 hours in the North China region, with a spatial resolution of 5 km, based on multiple numerical model precipitation~~

95 ~~forecasts. Compared to previous studies, this research offers the following advancements:~~ Short-term heavy precipitation is typically characterized by sudden onset, short duration, small spatial scale, and high localization. These features demand precipitation forecasts with finer temporal and spatial resolutions to meet operational needs. In this study, we employ a GAN-based model, GFRNet, to generate deterministic forecasts of 3-hour accumulated precipitation over the next 24 hours in North China, using multiple NWP model outputs as input and targeting a resolution of 5 km. Compared with previous research, this

100 study introduces the following key advancements:

- Focus on Severe Precipitation Events: ~~We place greater emphasis on forecasting severe precipitation events, adopting a threshold of $40mm/3hr$ for rainstorm classification, significantly higher than the $20mm/3hr$ or $5mm/hr$ thresholds used in prior research, thereby focusing on more intense precipitation events.~~ This study emphasizes the prediction of high-impact precipitation by adopting a more stringent threshold of 40 mm/3h for classifying rainstorms, in contrast to

105 the 20 mm/3h or 5 mm/h thresholds commonly used in prior work. This allows for better targeting of the most hazardous precipitation events.

- Application of GAN Strategy: ~~We apply the GAN strategy in developing the GFRNet model, which not only improves precipitation accuracy but also generates realistic predictions, effectively addressing the issue of blurriness that commonly plagues precipitation forecasts.~~ We implement a GAN-based approach in the GFRNet model, which not only enhances

110 precipitation prediction accuracy but also preserves fine-scale spatial structures. This effectively addresses the common issue of forecast blurriness and improves realism in predicted precipitation fields.

2 Data and method

2.1 Data

This study focuses on North China (35°N - 44.55°N, 112°E - 121.55°E), as illustrated in Figure 1. Administratively, this region

115 includes Beijing, Tianjin, Hebei, Shanxi, and the Inner Mongolia Autonomous Region, with the southeastern part encompassing Shandong and the Bohai Sea region. The target area features ~~a~~ complex topography, dominated by the Taihang Mountains, which extend from the southwest to the northeast. To the southeast lies the North China Plain, characterized by an average elevation below 400 meters. West of the Taihang Mountains is the Loess Plateau, and to the north is the Inner Mongolia Plateau, with elevations exceeding 800 meters and local peaks reaching up to 2000 meters.

120 This study utilizes CMA Multi-source merged Precipitation Analysis System (CMPAS) as the ground truth for precipitation fields. CMPAS is a comprehensive precipitation product developed by the National Meteorological Information Center of the China Meteorological Administration. It integrates ground automatic station data, satellite, and radar observations using methods such as Probability Density Function (PDF), Bayesian Model Averaging (BMA), Optimal Interpolation (OI) and Downscaling (DS)(Pan et al., 2018). CMPAS provides hourly temporal resolution ~~offers a temporal resolution of 1 hour~~ and a
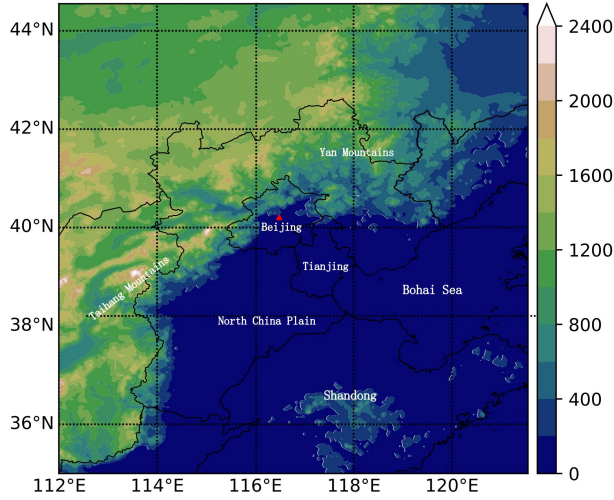
125 spatial resolution of 0.05° × 0.05°.

Figure 1. Topography distribution (shaded; in units of m) of HuaBei domain (35° - 45°N, 112° - 122°E.) . The vast area with an altitude of less than 400m in the middle and southeast of the figure is the North Chian Plain, which reaches the southern foot of Yanshan Mountain in the north, leans on Taihang Mountain in the west, and borders the Bohai Sea in the east. It includes Beijing (Red Triangle), Tianjin, Shandong, and most of Hebei.

For numerical models, considering the operational usage, model resolution, and performance, this study uses the precipitation forecast of the following three NWPs. The high-resolution global model forecast from the European Centre for Medium-Range Weather Forecasts (ECMWF), with a horizontal resolution of approximately 9 km in the China region and a temporal resolution of 3 hours; The mesoscale forecast from the East China Regional Numerical Center (CMA-SH9) (Zhang et al., 2021), with a horizontal resolution of 9 km and a temporal resolution of 1 hour; The high-resolution regional numerical forecast independently developed by the Numerical Prediction Center of the China Meteorological Administration (CMA-3KM) (Shen et al., 2020), with a horizontal spatial resolution of about 3 km and a temporal resolution of 1 hour. Forecasts are taken from the initial times of 00 UTC and 12 UTC, retaining a 24-hour forecast range. Spatially, numerical model forecasts are interpolated to a uniform grid of 0.05° × 0.05° using a bilinear interpolation algorithm, corresponding to a target area size of 192 × 192 grid points.

Based on the data described earlier, we performed a 3-hour accumulated precipitation (r3) forecast for the next 24 hours. Table 1 details the specific feature selection process, which includes five sources of features. ~~Let r3(T) denote the accumulated precipitation over the past 3 hours at time T, with the learning target being r3(T) from CMPAS~~. Let r3(T) denote the 3-hour accumulated precipitation at time T, where the learning target is the corresponding CMPAS r3(T) observation. The input features consist of r3(T) and r3(T-3) from ECMWF, CMA-SH9, and CMA-3KM. Given that precipitation formation, development, and movement are closely linked to topography and location, META features including elevation, latitude, and longitude are also incorporated into the model. The performance of numerical model forecasts varies depending on the forecast cycle and lead time. To account for this, temporal information such as forecast cycle and lead hour is encoded using trigonometric functions

and included as features in the deep DL model. The cycle values range from [0, 1], corresponding to the initial forecast times of 00 UTC and 12 UTC for the numerical models. For each cycle, only the forecast lead times at 3, 6, 9, 12, 15, 18, 21, and 24 hours are considered.

Table 1. ~~Data Sources and Features Used in Model~~ Data sources and features used in the model

	Source	Feature
	ECMWF	r3(T-3), r3(T)
	CMA-SH9	r3(T-3), r3(T)
Input	CMA-3KM	r3(T-3), r3(T)
	META	Elevation, Latitude, Longitude
	Time	Cos(cycle), Sin(cycle), Cos(lead hour), Sin(lead hour)
Label	CMPAS	r3(T)

Using the available data from 2019 to 2022, we divided the dataset into training, validation, and test sets. In the North China region, precipitation is predominantly concentrated in the summer months, ~~especially~~ particularly in July and August. Therefore, ~~we selected the period from July 10, 2021, to August 20, 2021, as the validation set, consisting of 637 samples. The period from June 15, 2022, to August 31, 2022, was designated as the test set, containing 1204 samples.The remaining data were allocated to the training set, resulting in 4645 samples.~~ the period from July 10 to August 20, 2021, was selected as the validation set, comprising 637 samples, while the period from June 15 to August 31, 2022, was designated as the test set, containing 1,204 samples. The remaining data were assigned to the training set, resulting in 4,645 samples. Since precipitation mainly occurs in the summer, with fewer events during other times of the year, it is crucial to apply reasonable sampling strategy for the training set. We aim to exclude non-precipitation samples and retain samples with a high proportion of precipitation areas or high precipitation intensity. The sampling rule is as follows: for a given sample, if the proportion of pixels with precipitation greater than threshold t exceeds r, the sample is retained; otherwise, it is discarded.

In this study, we set t=1 mm and r=2% ~~for sampling~~. The 1 mm precipitation threshold is low enough to capture the vast majority of precipitation events that have a real impact. At the same time, the sample proportion of 2% ensures the representativeness of the samples, so that the model can effectively learn and predict ~~important~~ crutial precipitation patterns ~~despite the~~ under limited computational resources. After applying these thresholds, the training set consisted of 2,885 samples. It is important to note that, to objectively reflect the model's forecasting capability in real-world scenarios, no ~~sample resampling~~ sampling or filtering was performed on the validation and test sets. see Table 2 for details.

We define valid precipitation samples as follows.

– Valid light rain samples: Samples where the proportion of pixels with precipitation $\geq$ 0.1mm exceeds 10%.

– Valid moderate rain samples: Samples where the proportion of pixels with precipitation $\geq$ 10mm exceeds 0.5%.

Table 2. Sample distribution across training, validation, and test sets.

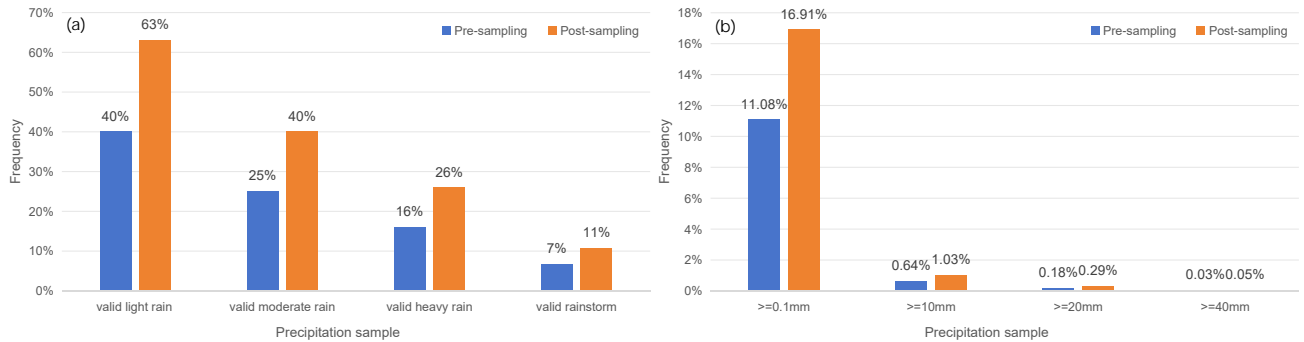| Dataset	Time Period	Samples	
		Pre-sampling	Post-sampling
Training set	2019-06-01 - 2019-10-10 2020-06-01 - 2020-10-10 2021-03-15 - 2021-07-09 2021-08-21 - 2021-10-10 2022-03-15 - 2022-06-14	4645	2885
Validation set	2021-07-10 - 2021-08-20	637	No sampling
Test set	2022-06-16 - 2022-08-31	1204	No sampling

Figure 2. Sample proportion distribution of precipitation with different intensity on training set before and after sampling (a)valid sample on image-level and (b) sample on pixel-level.

– Valid heavy rain samples: Samples where the proportion of pixels with precipitation $\geq$ 20mm exceeds 0.2%.

– Valid rainstorm samples: Samples where the proportion of pixels with precipitation $\geq$ 40mm exceeds 0.1%.

As shown in Figure 2a, the proportions of valid light rain samples and valid moderate rain samples increased from 40% and 25% to 60% and 40%, respectively. The proportions of valid heavy rain samples and valid rainstorm samples also reached 26% and 11%, respectively. This increase in the proportion of valid samples improved the stability and efficiency of model training.

When using loss functions like MSE or MAE to guide model updates, the loss is calculated at the pixel level. Therefore, although the image-level sampling strategy helps improve the efficiency of model learning, the distribution of precipitation within each pixel of the sampled image-level samples still exhibits a significant long-tail distribution (as shown in Figure 2b). This makes ~~learning from extreme precipitation events challenging~~ it challenging for the model to effectively learn from extreme precipitation events. To address this issue, we ~~have~~ designed a specialized weighted loss function, as ~~detailed~~ described in Section 2.2.

2.2 Method

2.2.1 Model

180   The core idea of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) is to use adversarial training to enable the Generator (G) to learn the distribution of real data and generate synthetic data that closely approximates real data. Simultaneously, the Discriminator (D) strives to improve its ability to distinguish between real data from the data set and data generated by the generator. In this study, we proposed a Generative Fusion Rain Net(GFRNet) for multi-NWP precipitation post-processing. As illustrated in Figure 3, GFRNet consists of two main components: the Generator and the Discriminator.

185   The core structure of the Generator in GFRNet was inspired by a U-Net with an encoder-decoder architecture(Ronneberger et al., 2015). The input to the model is a tensor of size 13×192×192 and the output is a tensor of size 1×192×192. The encoder comprises four Down-ConvBlocks, which gradually reduce the spatial dimensions of the feature maps while extracting deep feature information. The decoder, conversely, consists of four Up-ConvBlocks that progressively restore the spatial dimensions of the feature maps through upsampling operations. The specific sizes of the feature maps are illustrated in Figure 3a. Skip con-

190 nections are introduced between the encoder and decoder, connecting the output of a layer in the encoder directly to the input of the corresponding layer in the decoder. This helps better utilize the features extracted by the encoder. The activation function of the generator's final layer is set to ReLU(Agarap, 2019) for regression predictions. Each ConvBlock module consists of four parts:

- Convolution Operation: This transforms the size of the feature map, used for either upsampling or downsampling.

195   - Batch Normalization (BN) (Ioffe and Szegedy, 2015), ReLU, and Dropout(Srivastava et al., 2014) layers: These are used to accelerate the training process, improve model robustness, and prevent overfitting.

- Residual(He et al., 2015) module: This backbone consists of two convolutional layers with BN and dropout layer at the middle of it. The final output is obtained by adding the input data to the output of the second convolutional layer through skip connection.

200   - SE-Block: This is a channel attention module composed of two sub-modules: Squeeze and Excitation(Hu et al., 2019). The squeeze operation compresses the feature values of each channel via global pooling to obtain channel importance coefficients, and the excitation operation weights the feature map of each channel according to these coefficients.

  The Generator's U-Net-like structure can effectively capture the geographic and spatial dependencies of precipitation distribution. The residual structure in the ConvBlock can prevent gradient disappearance and explosion in deep-layer networks,

205 enhancing model performance and accelerating training. ~~In addition~~ Moreover, it improves the reuse and transmission of features. SE attention mechanisms helps the model focus on the feature channels that contribute significantly to the prediction of precipitation.

  Radford et al. (2016) significantly improved the training stability of GAN and the quality of generated images by introducing Deep Convolutional Network into GAN (DCGAN) structure. Inspired by the DCGAN, the main architecture of the
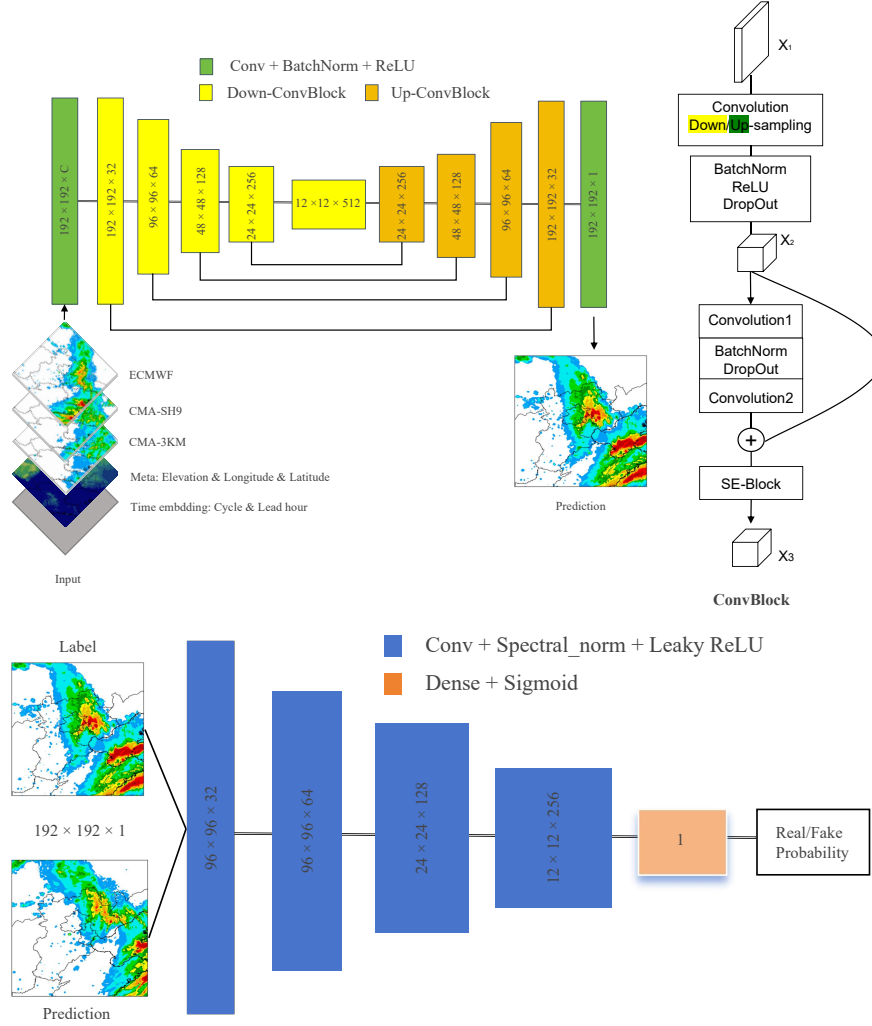
Figure 3. Model architecture (a)the generator of GFRNet model, also named FRNet and (b)Discriminator of GFRNet model.

210 our discriminator consists of four ConvBlocks that perform progressive spatial downsampling and channel expansion on the single input image of size 192×192×1, enabling richer semantic feature extraction. This is followed by a Dense layer and a Sigmoid activation, which outputs the probability that the input image is a real sample.. ~~These ConvBlocks perform four spatial downsamplings and channel dimension expansions on the single input image with size 192x192x1, extracting richer semantic information. Following this, a Dense layer and a Sigmoid layer are connected externally, outputting the probability of the image~~

215 ~~being a real sample.~~

9

2.2.2 Training

During the GAN training process, the generator and discriminator continuously compete and collaborate, driving mutual evolution. The generator aims to produce samples that resemble real data, while the discriminator receives both real and generated data as input and outputs a probability value indicating its confidence in the input being real. In this study, the optimization objectives for the discriminator and the generator are as follows:

$$\min_{\theta_D} E_{y,\hat{y}} \left[L_D \left(y, \hat{y}; \theta_D \right) \right] \tag{1}$$

$$\min_{\theta_G} E_{y,\hat{y},x} \left[L_G \left(y, \hat{y}, x; \theta_G \right) \right] \tag{2}$$

L_D and L_G represent the loss functions of the discriminator and generator, respectively. The parameters of the corresponding neural networks are denoted by θ_D and θ_G. The input to the generator and the predicted results are represented by x and $\hat{y}$, respectively, while y denotes the real labels. Wasserstein GANs (WGANs) (Arjovsky et al., 2017; Gulrajani et al., 2017) address the gradient vanishing problem commonly encountered in traditional GANs. Following the principles of WGAN, ~~this study uses~~ we adopt loss function with gradient penalty to optimize the discriminator. As shown in Equation 3, $D(y)$ and $D(G(\hat{y}|x))$ denote the scores assigned by the discriminator to real samples and samples generated by the generator, respectively. The latter part of the equation represents the gradient penalty term, where the weight γ is set to 10, the samples $\widetilde{y}$ are randomly weighted averages of the real label y and the generator's predicted $\hat{y}$, with ε being a randomly sampled value from a uniform distribution between 0 and 1.

$$L_D \left(y, \hat{y}; \theta_D \right) = 1 - D(y) + D(G(\hat{y} \mid x)) + \underbrace{\gamma \left(\left\| \nabla_{\tilde{x}} D(\widetilde{y} \mid x) \right\|_2 - 1 \right)^2}_{gradient\ penalty} \tag{3}$$

$$\widetilde{y} = \varepsilon y + (1 - \varepsilon) G(\hat{y} \mid x) \tag{4}$$

The loss function L_G for the generator consists of two components. The first part $D(G(\hat{y} \mid x))$ is the confidence score given by the discriminator, indicating how closely the generated images resemble real samples. We aim for this score to be as high as possible. The second part $L_{content}$ is the content loss, which is a weighted combination of Mean Squared Error (MSE) and Mean Absolute Error (MAE) loss functions. By setting the weight λ to 50, we ensure that the values of both loss components are on the same magnitude.

$$L_G \left(y, \hat{y}, x, \theta_G \right) = 1 - D(G(\hat{y} \mid x)) + \lambda L_{\text{content}} \tag{5}$$

$$L_{\text{content}} = L_{\text{wmse}} + L_{wmae} = \sum_{i=1}^{192}\sum_{j=1}^{192} w_{i,j}\left(y_{i,j} - \hat{y}_{i,j}\right)^2 + \sum_{i=1}^{192}\sum_{j=1}^{192} w_{i,j}||y_{i,j} - \hat{y}_{i,j}|| \tag{6}$$

$$w_{i,j} = \exp\left(ay_{i,j} + b\right) \tag{7}$$

In $L_{content}$, the MSE part emphasizes larger errors and provides a smoother gradient, while the MAE is less affected by outliers. Combining MSE and MAE helps balance large and small errors, enhancing the model's robustness and stability. Additionally, considering the long-tail distribution of r3 intensity as shown in Figure2, where significant precipitation events are ~~infrequent~~ rare but critical, it is crucial to assign higher loss weights to samples with strong precipitation intensity. ~~This approach prevents gradient explosion or vanishing issues during training and ensures the model effectively learns and predicts these rare, high-intensity precipitation events.~~ This strategy mitigates gradient vanishing or explosion and ensures the model learns to predict these rare, high-impact events effectively. As shown in Equation 7, ~~we found that when the loss weight is exponentially related to the precipitation amount with parameters a=4.3 and b=0.8, the model's performance is optimal.~~ We empirically found that using an exponential loss weighting function with parameters a=4.3 and b=0.8 yields optimal performance.

~~When training the model,~~ Both the generator and discriminator are optimized using the Adam optimizer Kingma and Ba (2017) with betas set to (0.9, 0.999) and a weight decay of 0.01. The learning rate follows a CosineAnnealingLR schedule Loshchilov and Hutter (2017), oscillating between 0.001 and 0 over a period of 20 epochs. During training, ~~it was~~ we observed that the discriminator~~'s ability~~ initially improved slowly, ~~therefore, it was necessary to slow down the training of the generator~~ necessitating a reduction in the generator's update frequency. Experimental results showed that updating the generator every 9 steps ~~resulted in more stable training for both the generator and the discriminator~~ stabilized training for both networks. ~~During the training process, the loss value on the validation set was used as the monitoring metric, and an early stopping strategy was employed.~~ Model training was monitored using the validation loss, and early stopping was employed. Training was halted if the validation loss ~~did not~~failed to decrease for 30 consecutive epochs. All evaluation results presented below are based on the model ~~parameters corresponding to~~ checkpoint with the lowest validation loss.

~~The number of parameters for the generator and discriminator is 4.46M and 0.72M,~~ The generator and discriminator contain 4.46M and 0.72M parameters respectively. ~~The training and prediction processes of GFRNet were executed using the NVIDIA(Compute Unified Device Architecture) CUDA library and NVIDIA Tesla graphics processing units (GPUs)~~ Training and inference were conducted using the NVIDIA CUDA library and Tesla GPUs . With a single NVIDIA A100 GPU, the training process completes in approximately 3 hours, and inference for 1,000 samples ~~can be completed in~~ takes just 2 minutes, ~~meeting the speed requirements for operational applications.~~ satisfying operational time constrains.

~~To thoroughly investigate the effectiveness of GFRNet, we conducted an experiment using a pure generator without the GAN strategy, referred to hereafter as FRNet. The content loss function, data, and training strategies for FRNet are consistent with those used for GFRNet.~~ To further evaluate GFRNet, we conducted an ablation study using only the generator, without

adversarial training, referred to as FRNet. The content loss, dataset, and training strategies for FRNet remain consistent with those used for GFRNet.

2.3 Evaluation metric

To evaluate the model's prediction results, we use the following metrics: Threat Score (TS), Probability of Detection (POD),
275 False Alarm Rate (FAR), and BIAS score. The specific definitions are as follows:

$$TS = \frac{h}{h+f+m} \tag{8}$$

$$POD = \frac{h}{h+m} \tag{9}$$

280 $$FAR = \frac{f}{h+f} \tag{10}$$

$$BIAS = \frac{h+f}{h+m} \tag{11}$$

The definition of h, f, m align with the confusion matrix shown in Table 3. The TS, POD, and FAR values range between 0 and 1. Higher TS and POD values and lower FAR values indicate better forecast performance. A BIAS value of 1 indicates an
285 unbiased forecast, while values between 0 and 1 indicate under-prediction, and values greater than 1 indicate over-prediction. In this study, thresholds of 0.1, 10, 20, and 40 mm, corresponding to light rain, moderate rain, heavy rain, and rainstorm, respectively, are used to comprehensively evaluate the model's performance.

Table 3. Confusion matrix to calculate metrics.True or False is determined by the chosen threshold

Confusion matrix		Observation	
		True	False
Prediction	True	Hit(h)	False alarm(f)
	False	Miss(m)	True negative(tn)

The metrics mentioned above are all measured by comparing individual pixel values. Even if the predicted rainfall structure and intensity match the actual conditions, a slight positional deviation in the predicted rainfall band from the observed
290 location can result in a high FAR and a lower POD, leading to a lower TS score, which cannot objectively reflect the true forecasting ability of the model. To address this issue, neighborhood spatial verification methods like the Fraction Skill Score (FSS) (Roberts and Lean, 2008) have been developed. FSS compares features within corresponding neighboring regions in

~~the forecast and observation fields. This method effectively evaluates the high-resolution model's capability to predict spatial structures, offering a more objective assessment of forecast quality. Additionally, FSS is straightforward to construct and is not~~
~~influenced by complex factors such as filtering thresholds or smoothing radius, resulting in consistent evaluation conclusions.~~
~~FSS has become a widely used spatial verification method and has been adopted by ECMWF as a standard precipitation~~
~~evaluation method to replace traditional precipitation skill scores. The FSS formula is defined as follows :~~ FSS evaluates forecast performance by comparing the fraction of grid points exceeding a certain threshold within a neighborhood in both forecast and observation fields. This approach enables a more objective assessment of high-resolution models' ability to capture spatial structures. Additionally, FSS is easy to implement and is not sensitive to parameters such as threshold filters or smoothing radii, which contributes to its consistent evaluation results. FSS is now widely used and has been adopted by ECMWF as a standard metric for precipitation evaluation, replacing many traditional skill scores. The FSS is derived from the Fractional Brier Score (FBS) and is calculated as follows:

$$FBS = \frac{1}{N} \sum_{i=1}^{N} (O_r - M_r)^2 \tag{12}$$

$$FSS = 1 - \frac{FBS}{\frac{1}{N} \left(\sum_{i=1}^{N} O_r{}^2 + \sum_{i=1}^{N} M_r{}^2 \right)} \tag{13}$$

Here, N is the total number of grid points within the evaluation domain, M_r and O_r represent the ratio of grid points exceeding a threshold to the total number of grid points within a given window size for the forecast and observation fields, respectively. First, we use a modified Brier score to compare the precipitation frequency between forecasts and observations, known as the Fraction Brier Score (FBS). Then, employing the variance skill score concept, we derive the Fraction Skill Score (FSS), which ranges from 0 to 1, where 0 indicates no match and 1 indicates a perfect match. ~~Typically, the FSS value increases as the neighborhood scale increases.~~ FSS typically increases with larger neighborhood sizes. From the definitions of FBS and BIAS, it can be observed that if the BIAS within the given window is significantly greater or less than 1, the FBS value increases, leading to a lower FSS score. This indicates that FSS penalizes both under-prediction ($BIAS < 1$) and over-prediction ($BIAS > 1$). To further assess the ~~accuracy of forecast images~~ quality of the predicted precipitation fields, we also use the Root Mean Square Error (RMSE) and the Multi-Scale Structural Similarity Index (MS-SSIM) (Wang et al., 2003). ~~MS-SSIM evaluates image similarity by considering brightness, contrast, and structure, providing a score between 0 and 1, with higher values indicating greater similarity.~~ RMSE quantifies overall deviation between predicted and observed values, while MS-SSIM evaluates image similarity in terms of luminance, contrast, and structural information. MS-SSIM values range from 0 to 1, with higher scores indicating better agreement between predicted and observed precipitation structures.

3 Results

The statistical evaluation results on the test set are given below.

3.1 Overall evaluation

Table 4 presents the evaluation metrics for GFRNet, FRNet, and NWPs, including pixel-wise TS, BIAS, FAR and POD scores for different rainfall thresholds, and spatial-wise FSS scores (window size=5). Additionally, RMSE and MS-SSIM scores for each model are also assessed (Table 5). In general, both GFRNet and FRNet outperform NWP in TS scores, with GFRNet achieving the most optimal BIAS score near 1 and the highest FSS scores. GFRNet ranks second in RMSE next to ECMWF and has the best performance in MS-SSIM, indicating superior prediction of spatial structure and intensity across various rainfall levels.

For light rain ($r3 \geq 0.1$mm), the TS and FSS scores across models are similar, with GFRNet and FRNet slightly ~~outperforming ahead of~~ NWPs. GFRNet has a BIAS of 0.78, ~~showing a~~indicating slight underprediction, while ECMWF's BIAS of 1.44 ~~indicates slight overprediction~~ reflects a tendency to overpredict, resulting in higher POD and FAR values. For moderate ($r3 \geq 10$mm) and heavy rain ($r3 \geq 20$mm), GFRNet and FRNet ~~significantly~~ substantially outperform NWPs in TS. GFRNet's TS score for moderate and heavy rain ~~improved by 28% and 34%~~ are 28% and 34% higher, respectively, ~~compared to~~ than those of CMA-3KM. ~~FRNet's high BIAS reflects overprediction, lowering its FSS score, even below CMA-3KM.~~ The high BIAS of FRNet suggests overprediction, which suppresses its FSS score, sometimes falling below that of CMA-3KM. In contrast, GFRNet achieves the highest FSS score with the lowest FAR, improved POD compared to NWPs, and a BIAS close to 1.

~~For storm rainfall ($r3 \geq 40$mm), GFRNet leads in FSS, with a TS score slightly lower than FRNet but superior to NWPs.~~ For storm-level rainfall ($r3 \geq 40$mm), GFRNet achieves the highest FSS, while its TS is slightly lower than FRNet's but better than those of NWPs. compared to CMA-3KM, GFRNet's FSS and TS are improved by 15% and 20%, respectively. GFRNet maintains the lowest FAR but has a lower POD than FRNet and CMA-3KM. ~~While both CMA-SH9 and FRNet have BIAS values above 1.8, indicating overprediction, ECMWF's BIAS of 0.2 shows significant underprediction. GFRNet's BIAS of 0.60 indicates a slight underprediction.~~BIAS values for CMA-SH9 and FRNet both exceed 1.8, indicating overestimation, whereas ECMWF severely underpredicts with a BIAS of 0.2. GFRNet's BIAS of 0.60 suggests a mild underestimation.

Among NWPs, ECMWF tends to over-predict light rain but significantly underpredicts heavy precipitation. For $r3 \geq 10$mm, CMA-SH9 shows the highest FAR and BIAS among NWPs, indicative of overforecasting. CMA-3KM demonstrates relatively better forecasting skills for moderate ~~, heavy~~ to severe rainfall ~~compared to ECMWF and CMA-SH9, with~~ showing higher TS and POD, a BIAS closer to 1, and the highest FSS score among NWPs.

GFRNet and ECMWF yield comparable RMSE, both around 2.2, while CMA-SH9 ~~having the highest~~ exhibits the largest RMSE of 3. ~~ECMWF's lowest RMSE results from generally weaker rainfall predictions, while GFRNet maintains the second-lowest RMSE with high prediction accuracy.~~ECMWF's low RMSE is largely due to its generally weaker rainfall intensity predictions, whereas GFRNet maintains low RMSE while achieving high accuracy. ~~The GFRNet MS-SSIM score of 0.763, the highest among the models, indicates the greatest realism in the forecast results. Although FRNet has a high TS score, its high BIAS for heavy rainfall and lack of spatial detail result in a lower SSIM than GFRNet.~~ GFRNet also attains the highest MS-SSIM score

of 0.763, indicating the most realistic forecasted precipitation fields. FRNet, despite its high TS, suffers from elevated BIAS in heavy rainfall and less accurate spatial patterns, leading to a lower SSIM compared to GFRNet.

Table 4. The evaluation results of ECMWF CMA-SH9 CMA-3KM FRNet and GFRNet for r3 prediction for next 24h(3-h interval).TS BIAS FAR POD and FSS are listed. Note: The best and second best score of each metric are shown in bold and underlined.

Forecast result	Model	TS	FAR	POD	BIAS	FSS
	ECMWF	0.405	0.511	**0.706**	1.444	0.693
	CMA-SH9	0.400	0.436	<u>0.578</u>	**1.024**	0.689
$r3 \geq 0.1mm$	CMA-3KM	0.389	0.409	0.532	<u>0.899</u>	**0.704**
	FRNet	**0.416**	<u>0.379</u>	0.557	0.896	<u>0.702</u>
	GFRNet	<u>0.406</u>	**0.343**	0.515	0.784	0.700
	ECMWF	0.155	0.704	0.2245	<u>0.829</u>	0.443
	CMA-SH9	0.128	0.790	0.246	1.174	0.376
$r3 \geq 10mm$	CMA-3KM	0.167	0.734	0.310	**1.167**	<u>0.469</u>
	FRNet	**0.216**	<u>0.725</u>	**0.501**	1.822	0.465
	GFRNet	<u>0.214</u>	**0.683**	<u>0.398</u>	1.254	**0.530**
	ECMWF	0.066	0.805	0.091	0.466	0.248
	CMA-SH9	0.075	0.882	0.171	1.458	0.270
$r3 \geq 20mm$	CMA-3KM	0.108	0.830	0.227	<u>1.333</u>	<u>0.363</u>
	FRNet	**0.147**	<u>0.804</u>	**0.373**	1.901	0.352
	GFRNet	<u>0.145</u>	**0.748**	<u>0.254</u>	**1.006**	**0.427**
	ECMWF	0.019	<u>0.887</u>	0.023	0.200	0.092
	CMA-SH9	0.031	0.953	0.086	1.850	0.151
$r3 \geq 40mm$	CMA-3KM	0.047	0.926	<u>0.117</u>	<u>1.588</u>	0.198
	FRNet	**0.077**	0.889	**0.201**	1.810	<u>0.215</u>
	GFRNet	<u>0.056</u>	**0.858**	0.085	**0.603**	**0.228**

Table 5. The RMSE and MS-SSIM of ECMWF CMA-SH9 CMA-3KM FRNet and GFRNet for r3 prediction for next 24h(3-h interval).Note: The best and second best score of each metric are shown in bold and underlined.

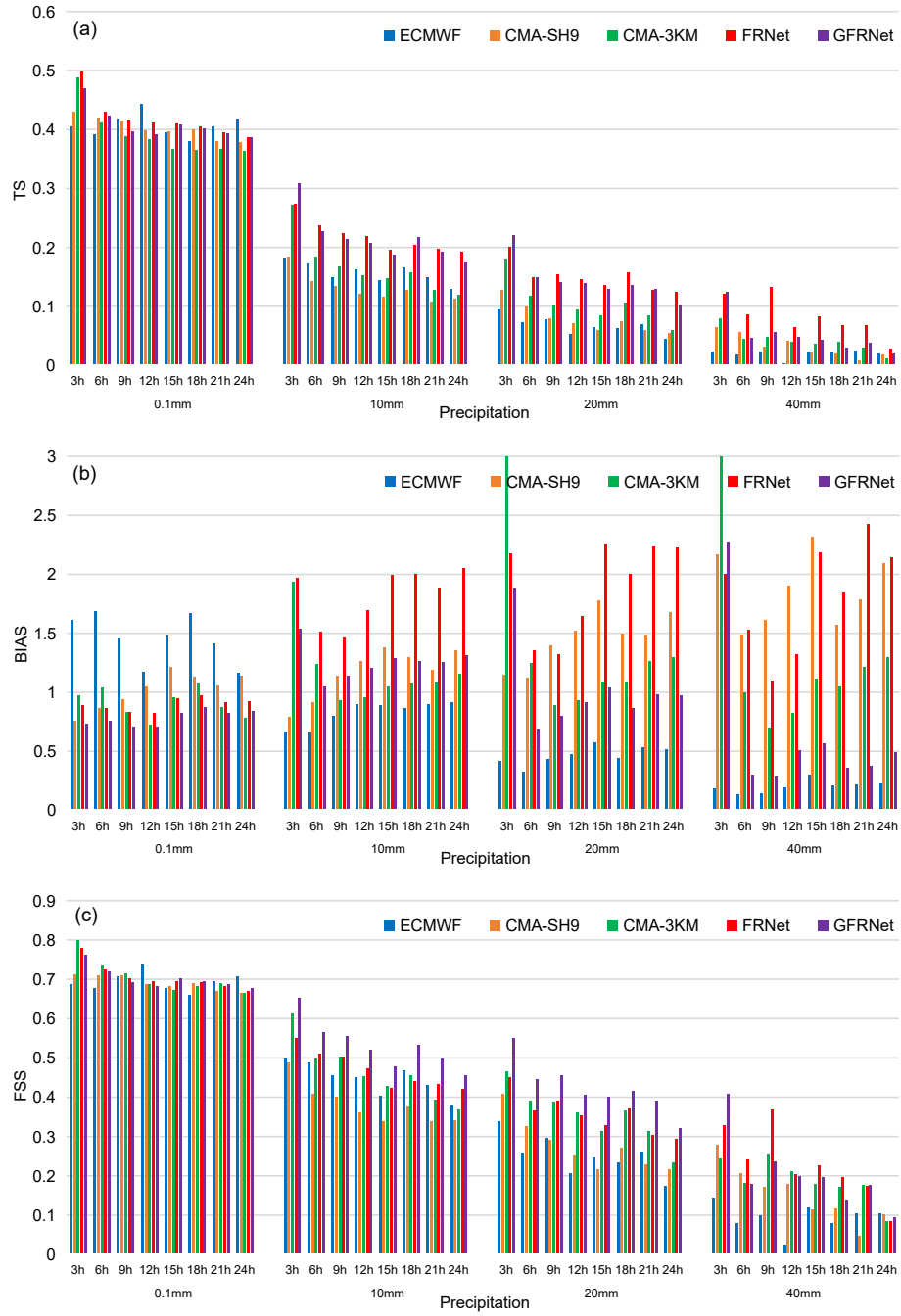| Model	RMSE	MS-SSIM
ECMWF	**2.208**	0.653
CMA-SH9	3.049	0.717
CMA-3KM	2.826	0.754
FRNet	2.459	<u>0.754</u>
GFRNet	<u>2.264</u>	**0.763**

Figure 4. TS BIAS and FSS temporal distribution of 3 - 24h precipitation forecasts from ECMWF, CMA-SH9, CMA-3KM, FRNet and GFRNet for 0.1, 10, 20, 40 mm(3h)-1. (a) TS score (b) BIAS score and (c) FSS score.

Figure 4 displays the TS, BIAS, and FSS scores at different lead times and rainfall thresholds. GFRNet generally outperforms NWPs across most lead times, although performance for all models declines with increasing lead time, as expected due to the limits of predictability.

For light rain, the TS and FSS scores of GFRNet and FRNet do not consistently surpass those of NWPs across all lead hours. CMA-3KM performs best at +3 h, while ECMWF leads at +12 h and +24 h. For moderate and heavy rain ~~predictions~~, GFRNet consistently ~~outperforms~~ surpasses NWPs in TS across all lead hours and ~~achieves they~~yields the highest FSS scores. BIAS mostly range between 0.6 and 1.5, remaining close to the ideal value. ~~For rainstorms~~ In storm events, GFRNet maintains lower BIAS values, around 0.5, at most lead hours, and while its TS and FSS are occasionally lower than thos of CMA-3KM, they are generally better than ECMWF and CMA-SH9.

For precipitation exceeding 10mm, FRNet's TS outperforms NWPs at all lead hours, but BIAS values are generally higher, mostly between 1.5 and 2.5, resulting in noticeable decreases in FSS scores, occasionally falling below CMA-3KM. Specifically, at lead time 3 h, both GFRNet and CMA-3KM exhibit notably elevated TS, FSS, and BIAS scores compared to other times. CMA-3KM's BIAS can exceed 3, likely due to overforecasting driven by its cloud initialization scheme. In contrast, GFRNet's BIAS at 3 h is generally below 2, indicating that GFRNet ~~can learn the characteristics of CMA-3KM and optimize corrections based on model biases.~~effectively learns from CMA-3KM's characteristics and adapts to correct its biases.

3.2 Spatial analysis

The spatial distribution of precipitation is closely related to topography. Figure 5 illustrates the spatial distribution of TS and BIAS scores ~~for various precipitation intensities across different models~~across different rainfall intensities for each model. Due to the limited test set of 1024 samples, calculating scores for each pixel would yield unrepresentative results. To address this, the 192 × 192 spatial domain was divided into 576 patches, each 8 × 8 in size, and metric scores were ~~calculated~~ computed for each patch to represent the spatial distribution.

For light rain, TS scores are relatively low in Inner Mongolia~~, while higher scores are observed~~but higher in Hebei, Shanxi, and Shandong. For moderate and heavy rain, peak TS values appear in northern Shanxi, the Bohai area, and central Inner Mongolia. Storm rainfall events, ~~primarily~~ which mainly occur east of the 500 m ~~contour line~~elevation contour, are concentrated in eastern Hebei, Bohai, and Shandong. Comparing GFRNet with NWPs, both GFRNet and FRNet effectively leverage the strengths of individual NWP. In regions where any NWP performs well, GFRNet generally achieves better performance, reflected in higher TS scores.

Regarding BIAS for light rain, all models ~~show BIAS < 0.6~~exhibit BIAS values below 0.6 in the Bohai area. Outside this region, ECMWF has BIAS above 1.5, while GFRNet, similar to CMA-3KM, maintains BIAS close to 1 cross most areas. For moderate and heavy rain, regions of high BIAS (BIAS > 2) in CMA-SH9 and CMA-3KM are primarily located west of the Taihang Mountains above 500 m elevation, indicating overprediction. For severe storm precipitation, high BIAS regions in CMA-SH9 and CMA-3KM are concentrated in Shanxi and the eastern Taihang Mountains. For precipitation above 10 mm, FRNet's high BIAS areas are concentrated east of the 500 m contour line, including most of Beijing, Tianjin, Hebei, Shandong, and Bohai, corresponding to its high TS areas. In contrast, GFRNet shows a relatively uniform BIAS distribution for moderate
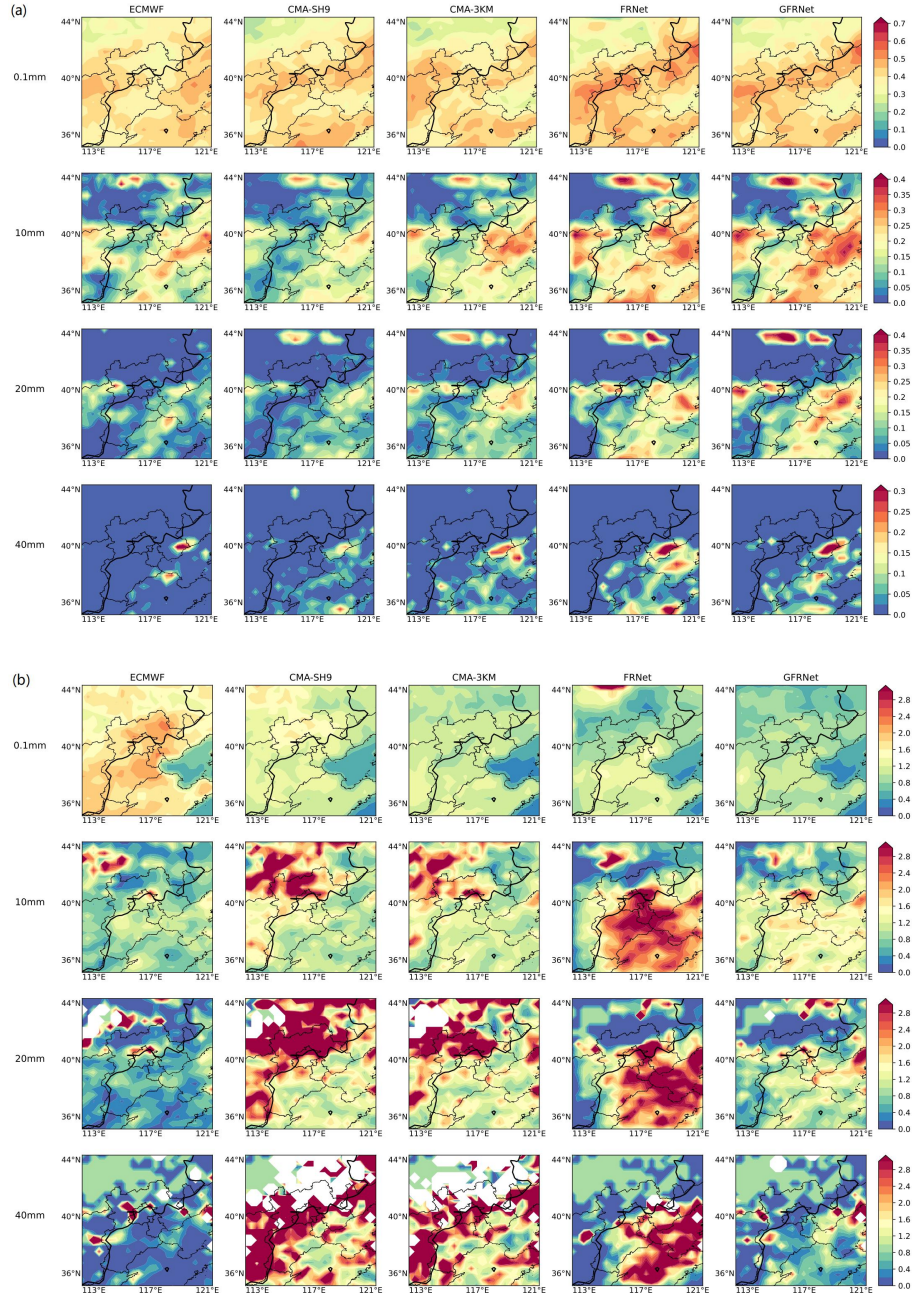
Figure 5. Spatial distribution of TS score and BIAS score on the test set. (a) TS score and (b) BIAS score. The solid black lines are 500m contours. The ECMWF, CMA-SH9, CMA-3KM, FRNet, and GFRNet models are represented from left to right columns, and 0.1 mm, 10 mm, 20 mm, and 40 mm(3h)-1 from top to bottom rows.
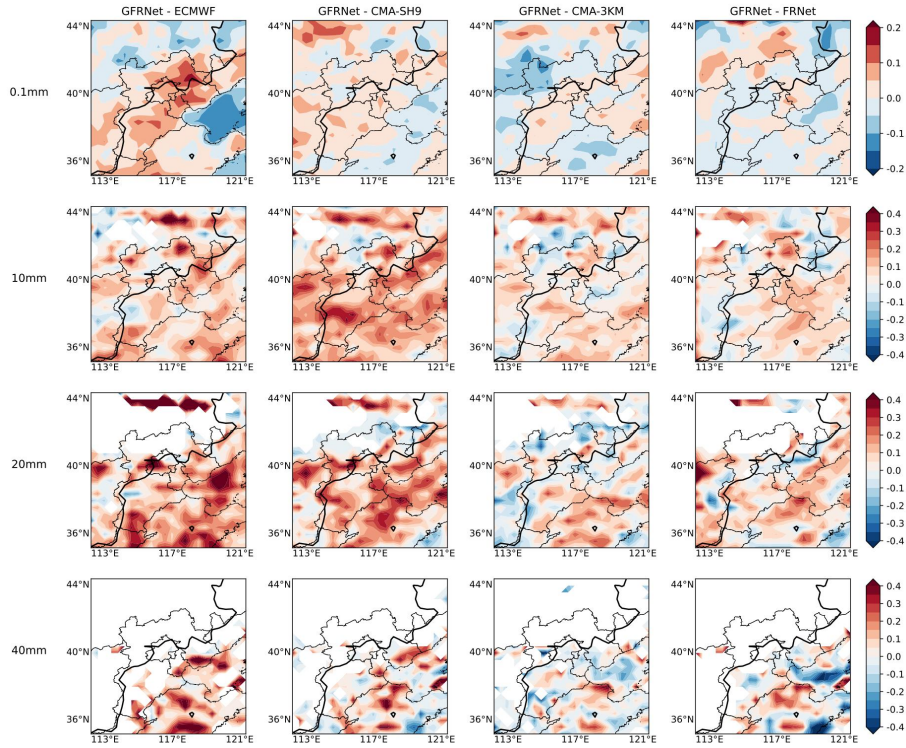
Figure 6. Spatial distribution of FSS gain for GFRNet on the test set, compared with ECMWF, CMA-SH9, CMA-3KM, and FRNet (from left to right columns) at precipitation thresholds of 0.1 mm, 10 mm, 20 mm, and 40 mm per 3 hours (from top to bottom rows). Black regions indicate terrain above 500 m. ~~The wihte area means both models get zero FSS score.~~ White areas denote grid cells where both models yield identical or zero FSS scores, either due to the absence of positive rainfall samples or limited predictive skill.

and heavy rain, with most BIAS values between 0.5 and 1.5. For severe storm rainfall, GFRNet's low BIAS areas align with the 500 m contour line, and BIAS intensity elsewhere is close to 1. This suggests that while NWPs and FRNet exhibit significant spatial BIAS heterogeneity, often leading to over- or under-prediction, GFRNet's precipitation spatial distribution is the most similar to the truth, with overall lower and more uniformly distributed spatial BIAS.

395   Figure 6 also presents the spatial distribution of GFRNet FSS gains compared to other models. For light rain, GFRNet's improvement zone extends from the southwest to the northeast, aligning with the direction and location of the 500 m contour line.In regions like Shandong and the Bohai Sea, however, GFRNet performs slightly weaker than NWP. For moderate and heavy rain, GFRNet shows improvement across nearly all regions, with the most significant gains in low-altitude or flat areas. The primary zones of heavy precipitation are Shandong and Bohai, where GFRNet demonstrates substantial FSS improve-

400   ments over ECMWF and CMA-SH9, and notable gains over CMA-3KM in central and northern Shandong. GFRNet not only ~~maintains~~ retains but also extends the ~~forecasting capabilities~~ skill of NWPs in regions where they perform well, showing

19

marked improvements. In areas where NWPs perform moderately, both GFRNet and FRNet show enhancements, though the ~~extent of improvement is~~ gains are somewhat limited.

~~Deep learning~~ DL approaches, can enhance forecasts of moderate and heavy precipitation in the western mountainous areas, eastern plains and ~~sea surfaces~~ coastal areas, and southern plains of the Beijing-Tianjin-Hebei region. However, FRNet's significant TS score improvements ~~come with an increase in~~are accompanied by the increased BIAS. In contrast, GFRNet achieves simultaneous improvements in both TS and BIAS, significantly reducing false alarms and effectively ~~preventing smoothing of spatial fields.~~preserving the spatial structure of the precipitation fields.

3.3 Case Study

3.3.1 Case1: 2022-07-05 00Z
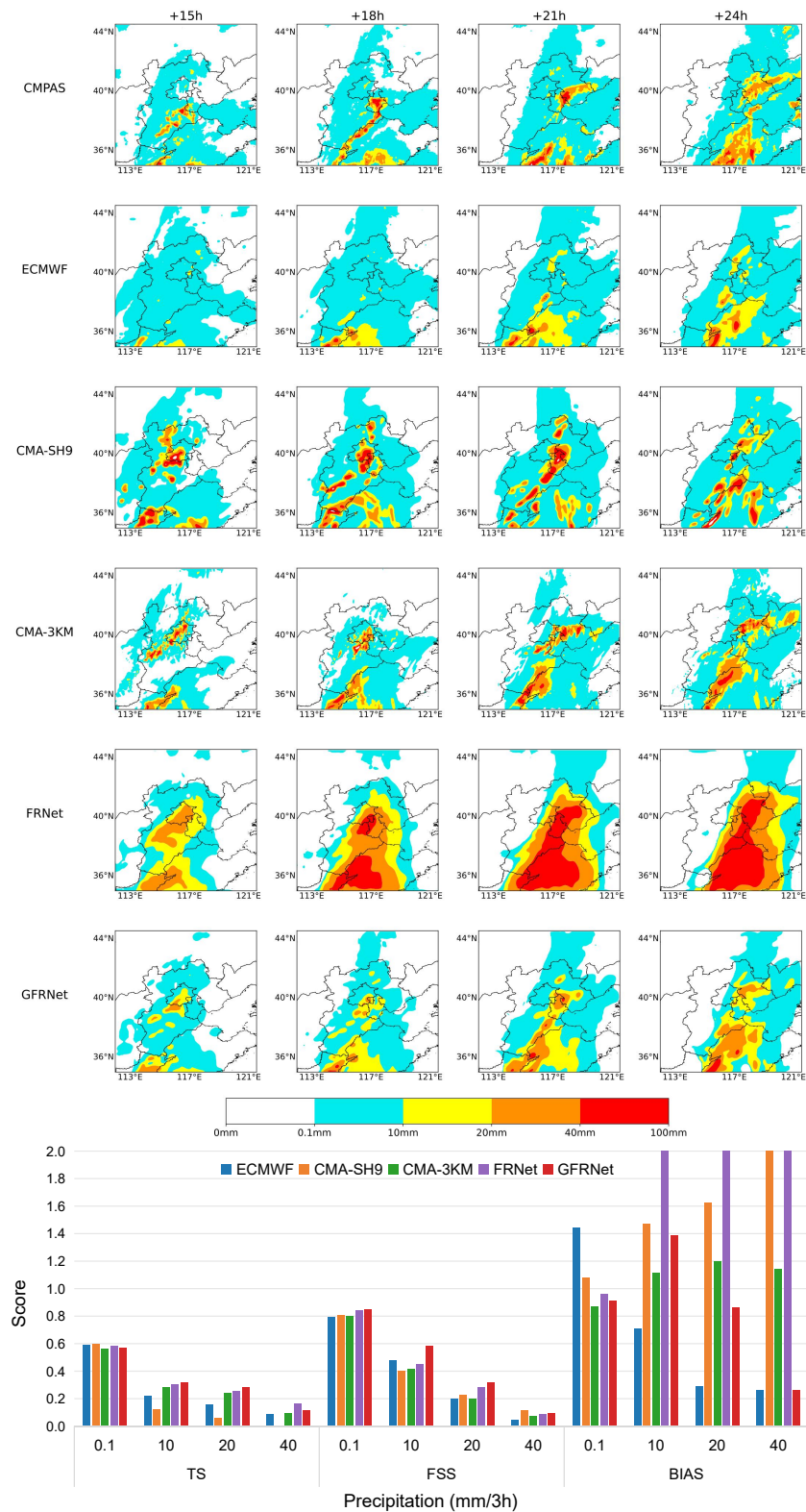
Figure 7. Precipitation forecasts of all models initiated at 0000 UTC in 5 July 2022 for the next 15h - 24h and the corresponding TS BIAS and FSS score of each models. OBS is the abbreviation CMPAS observation (same below).

On July 2nd, central and southern Hebei in North China experienced heavy rain ~~due to~~ under the influence of an upper-level trough. Meanwhile, southern China was affected by the typhoons "Chaba" and "Aere." By July 3rd, "Chaba" ~~had moved~~ reached approximately near 30°N, and by July 5th, it had weakened into an extratropical cyclone (figure omitted). The precipitation in North China was jointly driven by the upper trough and the weakening cyclone, resulting in significant rainfall cross the

415   southern and eastern protions ~~parts~~ of the region. This event ~~exemplifies~~ highlights the forecasting challenges ~~in North China related to typhoon influences~~ associated with typhoon influences in North China, particularly in predicting the development and intensity of convective systems winthin the warm sector ~~before~~ prior to the northward movement of the typhoon. ~~Additionally, there was considerable uncertainty in predicting the extent and intensity of rainfall caused by the merging of the upper trough with the weakening low-pressure system as it moved northeastward.~~ In addition, substantial uncertainty existed in forecasting

420   the extent and intensity of rainfall triggered by the interaction between the upper-level trough and the decaying low-pressure system as it migrated northeastward.

During this event (as shown on Figure 7), multiple small areas of heavy rainfall were observed at +15h, which ~~later coalesced~~ gradually merged into a long, narrow rainband alighned southwest to northeast by +18 h. ~~southwest-northeast oriented band of heavy rain by +18h.~~ The southernmost rain ~~cluster~~ cell developed and ~~moved~~ propagated northward. By +21 h and +24 h, the

425   heavy rainfall band ~~moved~~ shifted northeastward and split into two distinct clusters, with the southern cluster expanding in coverage. ECMWF's response to moderate and heavy rain was notably delayed, failing to predict the northern rain cluster, ~~However~~ though it performed reasonably well in ~~predicting the location of~~ locating the southern rainfall at +21 h and +24 h. Both the CMA-SH9 and CMA-3KM ~~correctly predicted the shapes and~~ successfully reporduced the spatial pattern and evolution of the two clusters, ~~though~~ albeit with slight positional deviations. ~~Additionally,~~ However CMA-9KM exhibited significant false

430   alarms and overpredicted rainfall intensity. The two ~~deep learning~~ DL models effectively captured the ~~trend of precipitation movement~~ overall movement of the precipitation. Although FRNet ~~accurately~~ precisely predicted the rain center and achieved the highest TS score, ~~it predicted overly smoothed results with high FAR and BIAS, limiting its operational value~~ its outputs were overly smoothed, leading to high FAR and BIAS values that limited its operational applicability. ~~GFRNet, on the other hand, integrated the location and intensity forecasting strengths of both ECMWF and CMA-3KM, enhancing the prediction~~

435   ~~of the location, intensity, and evolution of moderate to heavy rainfall.~~ In contrast, GFRNet effectively integrated the strengths of ECMWF and CMA-3KM in forecasting both location and intensity, yielding improved predictions of the spatial distribution, intensity, and evolution of moderate to heavy precipitation. Moreover, GFRNet provided clearer fine-scale precipitation structures, outperforming ~~numerical models~~ NWP models in both TS and FSS scores.
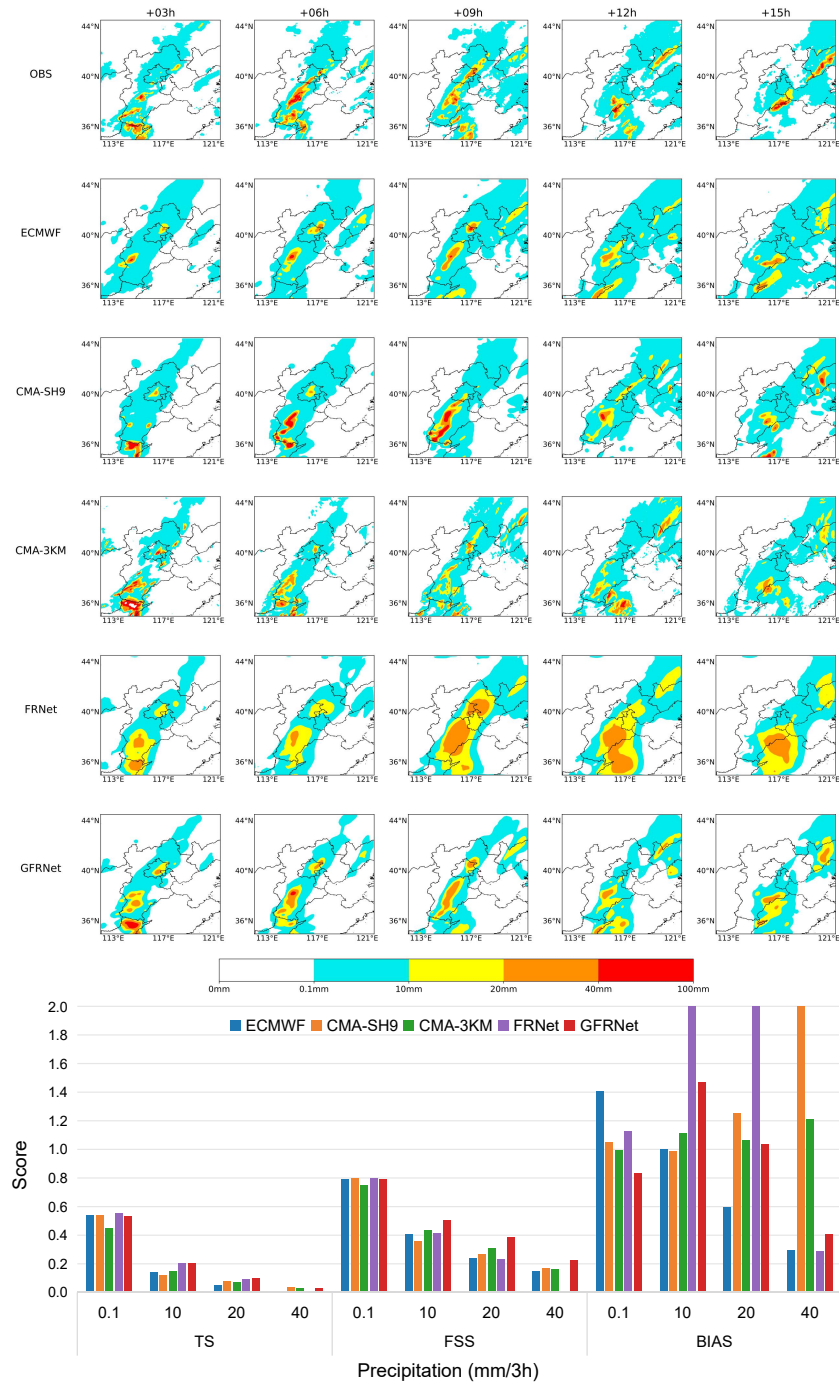
3.3.2 Case2: 2022-07-27 12Z

Figure 8. Precipitation forecasts of all models initiated at 1200 UTC on 27 July 2022 for the next 3h-15h and the corresponding TS BIAS and FSS of each models all showed below.

440   The heavy rainfall event on July 27th in North China was caused by the interaction between the ~~westerly belt cold air~~cold air from the westerly belt and the subtropical high. ~~NWPs perform well in capturing the general rainband but had significant biases~~NWPs generally captured the large-scale rainband well but exhibited substantial errors in predicting localized heavy rain. Among them, the ECMWF global model provided the most accurate ~~predicted the~~forecasts of the rainband's location, shape, and movement ~~of the rainband~~, ~~but its intensity forecast was too weak~~but significantly underestimated precipitation

445   intensity. The CMA-SH9 model predicted the highest ~~precipitation~~rainfall intensity but ~~had a significant delay in movement and overestimated the intensity~~displayed a marked delay in storm movement and a tendency to overestimate magnitude. The CMA-3KM model ~~did not perform well overall in predicting location and intensity but did accurately forecast some localized heavy rain centers~~performed poorly in forecasting both the location and intensity of precipitation, though it did correctly identify certain localized heavy rain centers.

450   As shown in Figure 8, at +03h, ~~there were~~Three small heavy rainfall areas ~~of heavy rain~~were observed in the southwest. The two northern centers moved northeastward and, by +06h and +09h, had formed a narrow southwest-northeast-oriented rainband on the eastern side of the Taihang Mountains, running paralleling to the mountain range. By +12 h, ~~this band broke into two centers again~~this band split once more into two distinct centers; the southern center weakened and dissipated by +15h as it moved eastward. At +03h, both GFRNet and CMA-3KM ~~accurately~~successfully predicted the three initial rain centers.

455   By +06 h and +09 h, CMA-3KM ~~significantly~~substantially underpredicted the linear convection, while ECMWF and CMA-SH9 performed better in ~~predicting~~capturing the rainband's location and intensity, respectively. ~~FRNet, despite improving its TS score through blurred predictions, lacked the capability to forecast the detailed structure and evolution of the heavy rain centers.~~FRNet, despite achieving a higher TS score through spatial smoothing, failed to reproduce the detailed structure and temporal evolution of the heavy rain centers. In contrast, GFRNet effectively combined the strengths of ECMWF and CMA-

460   SH9 in both position and intensity prediction, and successfully ~~predicting~~captured the evolution and fine-scale structure of the convective centers. GFRNet improved the TS score for moderate rain from 0.15 to 0.24 and for heavy rain from 0.1 to 0.14.

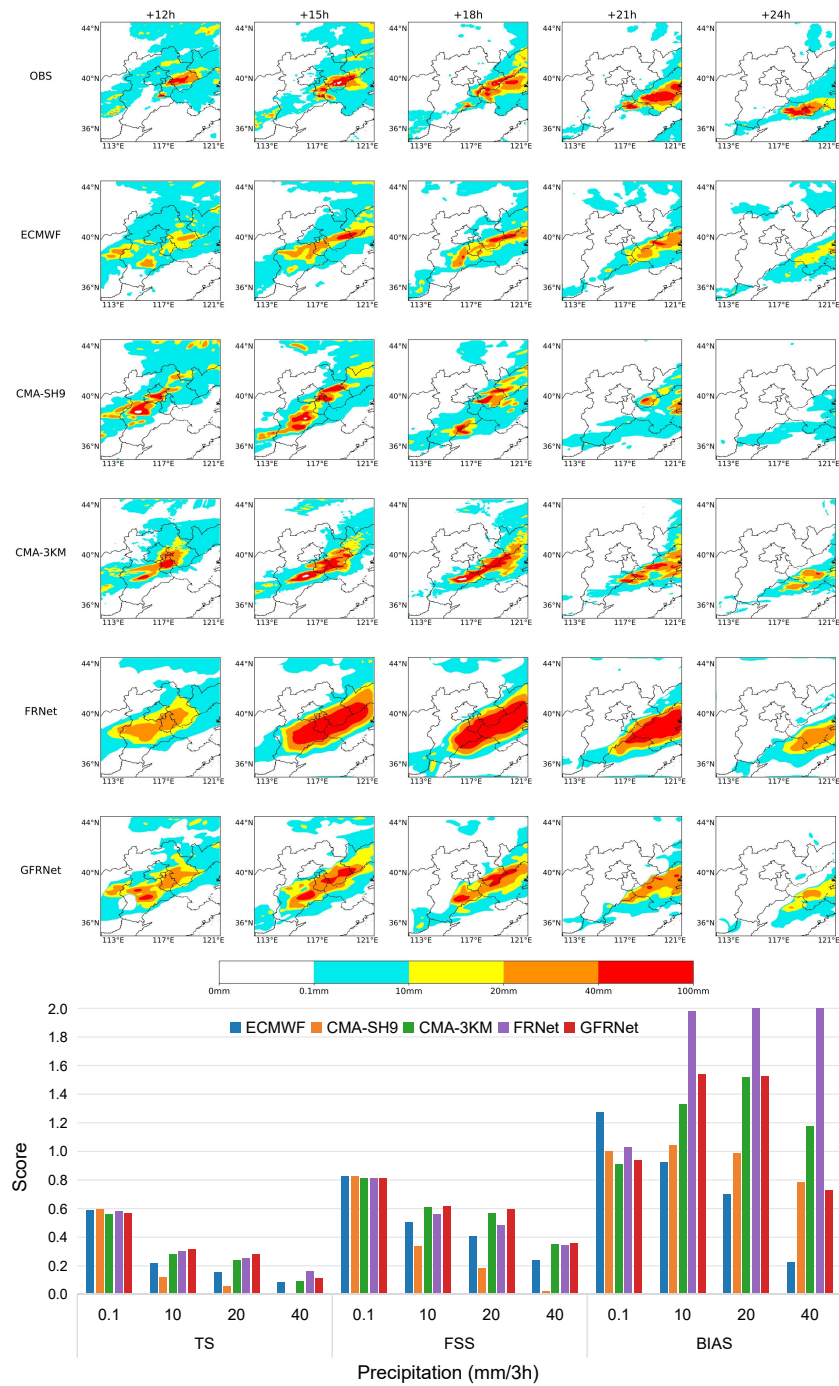### 3.3.3   Case3: 2022-08-18 00Z

Figure 9. Precipitation forecasts of all models initiated at 0000 UTC on 18 August 2022 for the next 12h-24h and the corresponding TS BIAS and FSS of each models all showed below.

On 18 August, a significant regional heavy rainfall event occurred in eastern North China, influenced by the northward extension of the subtropical high and a low-level jet stream. While the NWPs adequately forecasted the general precipitation area, they underestimated both the extent of the heavy rainfall centers and the intensity of extreme precipitation. As shown in Figure 9, the +12 h forecast indicated a strong precipitation center in the central-eastern ~~part of~~ North China, which subsequently extended and ~~developed~~ intensified southeastward, ~~with increasing intensity and area~~with both intensity and coverage increasing. The peak ~~was reached~~occurred at +21 h, and the heavy rainfall center persisted at +24 hours but ~~with a significantly smaller impact area~~its spatial extent had notably diminished. During this event, the ECMWF model accurately predicted the rainband's position and the ~~center of heavy~~ rain center, although ~~the intensity was~~ it slightly underestimated the intensity. The CMA-SH9 model predicted the precipitation center too far west and failed to forecast the rainfall center after +21 h. Both GFRNet and CMA-3KM provided consistent predictions regarding the location, intensity, and evolution of the precipitation center. Among these, the GFRNet's forecast for ~~the area~~ of moderate to heavy rainfall area was ~~larger and closer to the actual situation~~more extensive and closer to observation. The TS score for heavy rain increased from 0.23 to 0.28, and for storm-level rainfall, from 0.09 to 0.11, significantly higher than those of ECMWF and CMA-SH9, with BIAS values close to 1. In contrast, ~~while FRNet also had higher TS scores compared to NWP~~although FRNet achieved higher TS scores than NWPs, its BIAS values ~~were above~~exceeded 2, indicating ~~significant~~substantial overprediction.

GFRNet and FRNet demonstrated comparable TS performance for light rain to NWPs, with BIAS values closer to 1 than that of ECMWF. For moderate rain, heavy rain, and storm-level precipitation, both GFRNet and FRNet significantly outperformed NWPs in terms of TS scores. However, FRNet exhibited higher BIAS values and ~~tended towards more blurred~~produced overly smoothed predictions, lacking detailed precipitation ~~information, which was particularly evident in heavy precipitation.~~structures—especially for heavy precipitation. In comparison, GFRNet maintained BIAS values between 0.6 and 1.5, indicating no significant forecast bias, ~~and demonstrating the ability to forecast the formation, movement, and detailed~~and successfully captured the formation, movement, and fine-scale structure of rainbands and heavy precipitation centers. In terms of FSS scores, FRNet's ~~overly blurred predictions~~excessively smoothed outputs resulted in lower scores than CMA-3KM, whereas GFRNet consistently ~~showed the best spatial structure and morphology predictions for moderate rain and above~~produced the most accurate spatial morphology for moderate and heavier rainfall.

~~In analyzing the relationship between numerical models~~A comparative analysis of numerical models, FRNet, and GFRNet across the case studies~~, it is evident that~~reveals that NWP performance varies depending on the specific precipitation event. For instance, CMA-SH9 performed best in the second case, while CMA-3KM excelled in the third, followed by ECMWF. FRNet ~~employs a more coarse~~applied a more coarse-grained learning approach, primarily achieving ~~the lowest~~minimal loss ~~in the loss function through smooth predictions, sacrificing precipitation detail morphology.~~via smoothed predictions that sacrifice fine precipitation structures. ~~Although POD increased, BIAS and FAR also significantly increased.~~Although this leads to higher POD, it also results in significantly elevated BIAS and FAR. In contrast, GFRNet, using adversarial generation strategies, enables more refined model learning, as evidenced by: a. Avoiding the forecast defects of poorly performing ~~numerical models~~NWPs while dynamically learning ~~the advantages of~~from the strengths of better-performing models. b.

~~Ensuring accuracy while also predicting the detailed structure of precipitation~~Achieving accurate forecasts while preserving detailed precipitation structures, thereby enhancing ~~actual~~practical forecasting capabilities.

4 Discussion and conclusions

This study ~~used a GAN strategy to build~~utilized a GAN-based strategy to develop GFRNet for quantitative prediction of heavy rainfall in North China over the next 24 hours at 3-hour intervals. This model is based on forecast data from the global ECMWF model and the regional CMA-SH9 and CMA-3KM models. By ~~employing a reasonable sample sampling~~adopting a carefully designed sample selection strategy and a weight loss function to optimize the model training, GFRNet demonstrated superior performance in ~~forecasting precipitation~~predicting both intensity and location across all thresholds compared to NWPs on an independent validation ~~of the summer 2022 test set~~set for summer 2022, and the advantage is particularly pronounced for precipitation of 10 mm and above. Moreover, compared to the FRNet model, which does not use a generative adversarial ~~strategy~~framework, GFRNet significantly alleviates the common blurring issue prevalent in ~~deep learning~~ DL-based precipitation predictions.

FRNet primarily smooths predictions to minimize the content loss, sacrificing detailed precipitation structures. This ~~approach results in~~leads to a high BIAS score, with significant overprediction, and a lower FSS score compared to CMA-3KM. In contrast, GFRNet employs ~~co-evolution~~a co-evolutionary learning machanism between the generator and discriminator, leading to more refined learning. This allows the model to dynamically ~~integrate~~leverage the strengths of different numerical models, ~~ensuring accuracy while also predicting detailed precipitation structures.~~achieving both higher accuracy and more detailed precipitation structures.

In precipitation prediction tasks, conventional ~~deep learning~~ DL models often sacrifice ~~prediction clarity to enhance the~~forecast sharpness to improve TS score, ~~which significantly decreases their practical value.~~which substantially limits their operational utility. This study demonstrates that by adopting a Generative Adversarial Network (GAN) strategy ~~it is possible to improve accuracy without compromising the prediction of detailed precipitation structures, showing that these objectives~~can enhance forecast accuracy without compromising the representation of detailed precipitation structures, proving that these goals are not mutually exclusive.

Building on this foundation, future research will further explore the potential of generative models in precipitation forecasting. The adversarial architecture employed in this study is relatively simple, and future work can aim to improve both accuracy and sharpness by designing more advanced generative structures and incorporating conditional generation techniques.

One promising direction is the enhancement of forecast resolution. Current global and regional NWP systems are constrained to resolutions of 9 km and 3 km, respectively, due to computational limitations. However, generative models can be trained to upscale these forecasts to 1 km resolution using high-resolution observational datasets as supervision, addressing the growing demand for fine-scale precipitation forecasts.

Another important avenue is the generation of ensemble and probabilistic forecasts. Generative models can produce multiple forecast realizations through latent space sampling at a fraction of the computational cost of traditional ensemble NWP systems.

These ensemble forecasts provide valuable uncertainty quantification. However, key challenges remain in ensuring sufficient spread, diversity, and physical consistency among generated members.

Furthermore, the incorporation of physical constraints into generative models can enhance their ability to predict newly developed convection, especially in cases where none of the input NWP models capture the event. By integrating atmospheric variables such as temperature, humidity, pressure, and wind, the model can better learn complex physical relationships and improve its capability in forecasting convective initiation and evolution.

Finally, developing more comprehensive and operationally meaningful evaluation metrics is critical. While the Fractions Skill Score (FSS) improves upon TS by accounting for spatial displacement and intensity mismatch, it can still reward overly smoothed forecasts if TS is substantially improved. Such limitations can lead to misleading model assessments and hinder iterative improvements. Future evaluation frameworks should jointly consider pixel-level accuracy and spatial structural fidelity, penalize location and intensity errors, and emphasize the realism of forecasts to better reflect practical forecasting needs.

1. ~~Exploring Better Precipitation Forecast Generative Model Architectures: The current adversarial generative network used in this study is relatively preliminary. Future work can improve model accuracy and prediction clarity by designing more optimal adversarial generative network architectures and incorporating conditional generation ideas.~~

2. ~~Improving Effective Resolution of Precipitation Predictions Using Generative Models. Currently, due to limitations in computational resources and time, the highest resolution for global and regional precipitation forecasts of NWP is 9km and 3km, respectively, which cannot meet the growing demand for refined forecasts. Generative models can increase NWP forecasts to 1 km resolution by leveraging fine-grained 1km resolution precipitation reality products.~~

3. ~~Generating Ensemble and Probabilistic Forecasts with Generative Models. This approach can provide uncertainty information and improve forecast reliability, offering more scientific and practical forecast services to the public. While ensemble forecasting with numerical models is computationally expensive, generative models can generate multiple forecast ensembles through random sampling at a very low cost. However, ensuring reasonable dispersion, diversity, and complementarity among forecast members remains challenging.~~

4. ~~Incorporating physical guidance can enhance the model's ability to predict newly developed convection. Currently, GFRNet dynamically learns the strengths of precipitation forecasts from multiple numerical models, but it also performs poorly for precipitation that none of the numerical models can predict. By inputting information such as temperature, humidity, pressure, and wind speed into the model, it can help the model understand the non-linear and complex relationships between atmospheric physical states and precipitation, thus improving its ability to forecast newly developed convective precipitation.~~

~~Developing precipitation evaluation metrics that match forecasters' actual operational needs is also crucial. While FSS improves upon TS by considering spatial location and intensity deviations and penalizing blurry forecasts, it can still be misleading if the TS gains from blurred forecasts are sufficiently high. Unreasonable evaluation metrics hinder accurate assessment of model performance, impeding further model iteration. New metrics should comprehensively consider pixel-wise~~

565 *Code and data availability.* The gridded precipitation ground truth data and model forecast outputs used in this study are freely accessible at https://doi.org/10.57760/sciencedb.09821 (Zuliang and Qi, 2024).The code for training GFRNet and FRNet, as well as for evaluating model performance, is available at https://zenodo.org/records/14652556 (Fang and Zhong, 2025).

Appendix A: Ablation Study

A1 Training Ablation

570 Ablation experiments were conducted to assess the effects of two architectural components in GFRNet: the Squeeze-and-Excitation (SE) block and the weighted loss function. Specifically, the performance of GFRNet was compared with that of a variant without SE blocks (GFRNet_wo_SE) and another using standard MSE/MAE loss instead of the weighted loss (GFRNet_wo_WeightedLoss). The results are summarised in Table A1.

The SE blocks were found to have limited influence on light rain prediction, as reflected by the comparable TS scores of
575 GFRNet and GFRNet_wo_SE (0.406 vs. 0.408). However, for thresholds of 10 mm and above, GFRNet consistently outperformed the variant without SE blocks. For instance, at the 20 mm and 40 mm thresholds, the TS scores increased from 0.134 and 0.052 to 0.145 and 0.056, respectively. These results suggest that SE blocks play a vital role in capturing the structural details associated with heavier precipitation. In contrast, the use of standard loss functions led to improved TS scores for light rain (0.431 vs. 0.406), indicating better performance in this regime. Nevertheless, for higher thresholds, the weighted loss func-
580 tion significantly enhanced model accuracy. At the 20 mm and 40 mm thresholds, the TS scores of GFRNet_wo_WeightedLoss dropped to 0.115 and 0.028, compared to 0.145 and 0.056 for GFRNet. This demonstrates the effectiveness of the weighted loss in improving sensitivity to moderate and heavy rainfall.

Table A1. TS scores for different rain thresholds in blocks ablation experiments.Note: The best score is indicated in bold and the second-best score is underlined

Model/Threshold	0.1 mm	10 mm	20 mm	40 mm
GFRNet	0.406	**0.214**	**0.145**	**0.056**
GFRNet_wo_SE	<u>0.408</u>	<u>0.195</u>	<u>0.134</u>	<u>0.052</u>
GFRNet_wo_WeightedLoss	**0.431**	0.191	0.115	0.028

In summary, both SE blocks and the weighted loss function are essential to GFRNet's performance in forecasting moderate to heavy precipitation. The SE blocks enhance spatial feature representation, while the weighted loss improves model focus on

585 high-impact events. These findings confirm the utility of the proposed components in improving the robustness and accuracy of precipitation forecasts.

A2 Source Ablation

We conducted a series of ablation experiments to systematically evaluate the contribution of each input source to the precipitation forecasting performance of GFRNet. The influence of each input was quantified using a Relative Importance Score (RIS), 590 defined as:

$$\text{Relative Importance}(x) = \frac{\text{TS(GFRNet)} - \text{TS(GFRNet\_wo\_x)}}{\text{TS(GFRNet)}} \qquad (A1)$$

Table A2. TS scores for different rain thresholds in source ablation experiments

Model/Threshold	0.1 mm	10 mm	20 mm	40 mm
wo_ECMWF	0.397	0.186	0.127	0.053
wo_CMA-SH9	0.406	0.201	<u>0.137</u>	<u>0.055</u>
wo_CMA-3KM	0.405	0.183	0.124	0.051
wo_META	**0.418**	<u>0.203</u>	0.134	0.049
wo_Time	<u>0.410</u>	0.198	0.126	0.050
GFRNet	0.406	**0.214**	**0.145**	**0.056**

Table A3. RIS of each source on different rain threshold

Model/Threshold	0.1 mm	10 mm	20 mm	40 mm
ECMWF	**2.22%**	<u>12.93%</u>	12.41%	5.36%
CMA-SH9	0%	6.07%	5.52%	1.79%
CMA-3KM	<u>0.25%</u>	**14.34%**	**14.14%**	8.93%
META	-2.96%	5.11%	7.54%	**12.50%**
Time	-0.99%	7.48%	<u>13.10%</u>	<u>10.71%</u>

The TS scores of the ablation experiments are presented in Table A2. The results indicate that, even with the removal of any single input, GFRNet consistently outperforms the three NWP baselines in forecasting moderate, heavy, and storm precipitation. This demonstrates the model's robustness and its capacity to produce reliable corrections even when certain data 595 sources are unavailable.

Further analysis of the RIS values (Table A3) shows that, except for META and temporal features—which exhibit a minor negative effect on light rain forecasts—all inputs contribute positively across precipitation categories. The ECMWF input

is particularly beneficial for moderate and heavy rainfall, though it plays a lesser role in light and storm precipitation forecasts. This is consistent with its status as a global high-resolution model with advanced physical parameterizations (e.g. cloud microphysics and boundary-layer schemes), which enhances its skill in simulating mesoscale precipitation processes.

The CMA-3KM input yields substantial improvements across moderate, heavy, and storm precipitation forecasts, with particularly strong impact on moderate and heavy rain. As a high-resolution regional model, CMA-3KM is capable of resolving finer-scale convective structures and local precipitation evolution, thereby enhancing forecast accuracy in these regimes. In contrast, CMA-SH9 contributes modestly to moderate and heavy rainfall forecasts, but its impact on light and storm precipitation is limited—likely due to its lower spatial resolution and less detailed physical process representations.

META and temporal features improve forecasts for moderate to storm precipitation but slightly degrade performance for light rainfall, possibly due to increased noise. Heavier precipitation events tend to exhibit clearer spatial patterns and more distinct temporal evolution, which can be effectively leveraged by topographic and temporal features.

Overall, by integrating multiple NWP model outputs and auxiliary features, GFRNet substantially improves the accuracy and resolution of precipitation forecasts. The ablation results highlight the model's effectiveness in forecasting moderate to extreme precipitation and demonstrate its robustness to missing input sources.

References

Agarap, A. F.: Deep Learning using Rectified Linear Units (ReLU), https://arxiv.org/abs/1803.08375, 2019.

Arjovsky, M., Chintala, S., and Bottou, L.: Wasserstein GAN, https://arxiv.org/abs/1701.07875, 2017.

Boeing, G.: Visual Analysis of Nonlinear Dynamical Systems: Chaos, Fractals, Self-Similarity and the Limits of Prediction, Systems, 4, 37, https://doi.org/10.3390/systems4040037, 2016.

Chen, P. J., Feng, Y. R., Meng, W. G., Wen, Q. S., Pan, N., and Dai, G. F.: A correction method of hourly precipitation forecast based on convolutional neural network, Meteor Mon, 47, 60–70, https://doi.org/10.7519/j.issn.1000-0526.2021.01.006, 2021.

Chen, Y., Huang, G., Wang, Y., Tao, W., Tian, Q., Yang, K., Zheng, J., and He, H.: Improving the heavy rainfall forecasting using a weighted deep learning model, https://doi.org/10.3389/fenvs.2023.1116672, 2023.

Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C., Gazen, C., Carver, R., Andrychowicz, M., Hickey, J., Bell, A., and Kalchbrenner, N.: Deep Learning for Twelve Hour Precipitation Forecasts, Nature Communications, 13, 5145, https://doi.org/10.1038/s41467-022-32483-x, 2022.

Fang, Z. and Zhong, Q.: Improving the fine structure of intense rainfall forecast by a designed adversarial generation network, https://doi.org/10.5281/zenodo.14652556, 2025.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative Adversarial Networks, https://arxiv.org/abs/1406.2661, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A.: Improved Training of Wasserstein GANs, https://arxiv.org/abs/1704.00028, 2017.

Harris, L., McRae, A. T. T., Chantry, M., Dueben, P. D., and Palmer, T. N.: A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts, Journal of Advances in Modeling Earth Systems, 14, e2022MS003 120, https://doi.org/10.1029/2022MS003120, 2022.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, https://arxiv.org/abs/1512.03385, 2015.

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E.: Squeeze-and-Excitation Networks, https://arxiv.org/abs/1709.01507, 2019.

Ioffe, S. and Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, https://arxiv.org/abs/1502.03167, 2015.

Kim, T., Ho, N., Kim, D., and Yun, S.-Y.: Benchmark Dataset for Precipitation Forecasting by Post-Processing the Numerical Weather Prediction, https://arxiv.org/abs/2206.15241, 2022.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, https://arxiv.org/abs/1412.6980, 2017.

Leinonen, J., Nerini, D., and Berne, A.: Stochastic Super-Resolution for Downscaling Time-Evolving Atmospheric Fields With a Generative Adversarial Network, IEEE Transactions on Geoscience and Remote Sensing, 59, 7211–7223, https://doi.org/10.1109/TGRS.2020.3032790, 2021.

Loshchilov, I. and Hutter, F.: SGDR: Stochastic Gradient Descent with Warm Restarts, https://arxiv.org/abs/1608.03983, 2017.

Pan, Y., Gu, J., Yu, J., Shen, Y., Shi, C., and Zhou, Z.: Test of merging methods for multi-source observed precipitation products at high resolution over China, Acta Meteorologica Sinica, 76, 755–766, https://doi.org/10.11676/qxxb2018.034, 2018.

Price, I. and Rasp, S.: Increasing the accuracy and resolution of precipitation forecasts using deep generative models, https://arxiv.org/abs/2203.12297, 2022.

655 Radford, A., Metz, L., and Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, https://arxiv.org/abs/1511.06434, 2016.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., and Mohamed, S.: Skillful Precipitation Nowcasting Using Deep Generative Models of Radar, Nature, 597, 672–677, https://doi.org/10.1038/s41586-021-03854-z,
660 2021.

Roberts, N. M. and Lean, H. W.: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, Monthly Weather Review, 136, 78–97, https://doi.org/10.1175/2007MWR2123.1, 2008.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, https://arxiv.org/abs/1505.04597, 2015.

665 Shen, X., Wang, J., Li, Z., Chne, D., and Gong, J.: China's independent and innovation development of numerical weather prediction, Acta Meteorologica Sinica, 78, 451–476, https://doi.org/10.11676/qxxb2020.030, 2020.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., kin Wong, W., and chun Woo, W.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, https://arxiv.org/abs/1506.04214, 2015.

Singh, A. K., Albert, A., and White, B.: Downscaling Numerical Weather Models with GANs, https://api.semanticscholar.org/CorpusID:
670 226785468, 2019.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, The Journal of Machine Learning Research, 15, 1929–1958, 2014.

Sun, D., Huang, W., Yang, Z., Luo, Y., Luo, J., Wright, J. S., Fu, H., and Wang, B.: Deep Learning Improves GFS Wintertime Precipitation Forecast Over Southeastern China, Geophysical Research Letters, 50, e2023GL104 406,
675 https://doi.org/https://doi.org/10.1029/2023GL104406, e2023GL104406 2023GL104406, 2023.

Sun, J., Xue, M., Wilson, J. W., Zawadzki, I., Ballard, S. P., Onvlee-Hooimeyer, J., Joe, P., Barker, D. M., Li, P.-W., Golding, B., Xu, M., and Pinto, J.: Use of NWP for Nowcasting Convective Precipitation: Recent Progress and Challenges, Bulletin of the American Meteorological Society, 95, 409–426, https://doi.org/10.1175/BAMS-D-11-00263.1, 2014.

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., Agrawal, S., Hickey, J., and Kalchbrenner, N.: MetNet: A
680 Neural Weather Model for Precipitation Forecasting, https://arxiv.org/abs/2003.12140, 2020.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., and Tang, X.: ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks, https://arxiv.org/abs/1809.00219, 2018a.

Wang, Y., Gao, Z., Long, M., Wang, J., and Yu, P. S.: PredRNN++: Towards A Resolution of the Deep-in-Time Dilemma in Spatiotemporal Predictive Learning, https://arxiv.org/abs/1804.06300, 2018b.

685 Wang, Z., Simoncelli, E., and Bovik, A.: Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2, pp. 1398–1402 Vol.2, https://doi.org/10.1109/ACSSC.2003.1292216, 2003.

Yang, X., Dai, K., and Zhu, Y.: Progress and challenges of deep learning techniques in intelligent grid weather forecast, Acta Meteorologica Sinica, 80, 649–667, https://doi.org/10.11676/qxxb2022.051, 2022.

Zhang, C.-J., Zeng, J., Wang, H.-Y., Ma, L.-M., and Chu, H.: Correction Model for Rainfall Forecasts Using the LSTM with Multiple
690 Meteorological Factors, Meteorological Applications, 27, e1852, https://doi.org/10.1002/met.1852, 2020.

Zhang, X., Yang, Y., Chen, B., and Huang, W.: Operational Precipitation Forecast Over China Using the Weather Research and Forecasting (WRF) Model at a Gray-Zone Resolution: Impact of Convection Parameterization, Weather and Forecasting, https://doi.org/10.1175/WAF-D-20-0210.1, 2021.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., and Wang, J.: Skilful Nowcasting of Extreme Precipitation with NowcastNet, Nature, pp. 1–7, https://doi.org/10.1038/s41586-023-06184-4, 2023.

Zhou, K., Sun, J., Zheng, Y., and Zhang, Y.: Quantitative Precipitation Forecast Experiment Based on Basic NWP Variables Using Deep Learning, ADVANCES IN ATMOSPHERIC SCIENCES, 39, 1472–1486, https://doi.org/10.1007/s00376-021-1207-7, 2022.

Zuliang, F. and Qi, Z.: Precipitaion observation and forecast in North China in 2022 by numerical model and deep learning model, https://doi.org/10.57760/sciencedb.09821, 2024.

695