remote sensing

MDPI

Article

MBFE-UNet: A Multi-Branch Feature Extraction UNet with Temporal Cross Attention for Radar Echo Extrapolation

Huantong Geng [1,2], Han Zhao [1,*], Zhanpeng Shi [1], Fangli Wu [3], Liangchao Geng [4] and Kefei Ma [3]

[1] School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; htgeng@nuist.edu.cn (H.G.); 202212200009@nuist.edu.cn (Z.S.)
[2] China Meteorological Administration Radar Meteorology Key Laboratory, Nanjing 210023, China
[3] School of Software, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202212210021@nuist.edu.cn (F.W.); 202312210045@nuist.edu.cn (K.M.)
[4] School of Atmospheric Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202311010012@nuist.edu.cn
* Correspondence: 202312490563@nuist.edu.cn

Abstract: Radar echo extrapolation is a critical technique for short-term weather forecasting. Timely warnings of severe convective weather events can be provided according to the extrapolated images. However, traditional echo extrapolation methods fail to fully utilize historical radar echo data, resulting in limited accuracy for future radar echo prediction. Existing deep learning echo extrapolation methods often face issues such as high-threshold echo attenuation and blurring distortion. In this paper, we propose a UNet-based multi-branch feature extraction model named MBFE-UNet for radar echo extrapolation to mitigate these issues. We design a Multi-Branch Feature Extraction Block, which extracts spatiotemporal features of radar echo data from various perspectives. Additionally, we introduce a Temporal Cross Attention Fusion Unit to model the temporal correlation between features from different network layers, which helps the model to better capture the temporal evolution patterns of radar echoes. Experimental results indicate that, compared to the Transformer-based Rainformer, the MBFE-UNet achieves an average increase of 4.8% in the critical success index (CSI), 5.5% in the probability of detection (POD), and 3.8% in the Heidke skill score (HSS).

Keywords: radar echo extrapolation; deep learning; UNet; multi-branch

Citation: Geng, H.; Zhao, H.; Shi, Z.; Wu, F.; Geng, L.; Ma, K. MBFE-UNet: A Multi-Branch Feature Extraction UNet with Temporal Cross Attention for Radar Echo Extrapolation. *Remote Sens.* **2024**, *16*, 3956. https://doi.org/10.3390/rs16213956

Academic Editor: Marios Anagnostou

Received: 31 July 2024
Revised: 19 October 2024
Accepted: 21 October 2024
Published: 24 October 2024

1. Introduction

Radar echo extrapolation refers to using historical echo observations to predict future echoes, which provides highly accurate data support for short-term weather forecasting and plays a critical role in predicting extreme weather events such as heavy precipitation, thunderstorms, and hail. Accurate extrapolation enhances the timeliness and reliability of weather forecasting and significantly impacts disaster prevention and mitigation, public safety, and the planning of economic activities. Therefore, in-depth research and optimization of radar echo extrapolation methods are crucial to modern meteorology and disaster management practices.

Traditional radar echo extrapolation methods primarily include centroid tracking [1–3], cross-correlation [4–6], and optical flow [7–10]. Centroid tracking predicts the short-term movement trend by calculating the centroid changes of echo cells in two consecutive images. However, this method overlooks the changes in echo morphology, leading to a significant increase in tracking failures in complex weather systems. The cross-correlation method divides radar echo images into multiple two-dimensional areas, determining the evolution trend of echoes by calculating the correlation coefficients of echoes in different regions at different times, thereby achieving the goal of tracking and extrapolating echoes. However, this method frequently encounters low correlation during strong convective weather, resulting in reduced accuracy in extrapolation. The optical flow method was first applied in

computer image processing to determine the motion characteristics of objects based on their spatiotemporal features and correlations of image sequences. This method necessitates that image sequences adhere to the assumption of brightness constancy, thereby enabling the calculation of the movement speed and direction of each pixel. However, the actual evolution of radar echoes is highly complex. When adjacent images undergo significant changes, the optical flow method yields substantial errors in extrapolation. In addition, the above traditional method fails to effectively utilize many historical observations of radar echo sequences. Hence, the forecasting timeliness and accuracy are far from meeting the forecasting requirements.

With the rapid advancement of deep learning technology, radar echo extrapolation has been increasingly regarded as a spatiotemporal sequence prediction problem [11–14]. As a result, various recurrent neural networks (RNNs) have been developed to enhance the precision and efficacy of this technique. For instance, the ConvLSTM model proposed by Shi et al. [15] opens the application of deep learning in radar echo extrapolation. By introducing the convolutional neural network (CNN) into the long and short-term memory (LSTM) network, this model reduces the number of parameters and better captures spatial and temporal dynamic information. However, this ConvLSTM places greater emphasis on temporal information and ignores the utilization of deep spatial features. To address this limitation, Wang et al. [16] suggested the PredRNN model, which is the first attempt to separate the spatial and temporal memory units. The model's unique Z-shape memory flow enables spatial information to propagate across all layers of the model. Subsequently, due to the issue of saturation in the forget gate of PredRNN, Wang et al. [17] put forward the Memory in Memory (MIM) model, which improves the forget gate using MIM-S and MIM-N modules that separately learn stationary and non-stationary features in spatiotemporal dynamics. However, their more complex model architecture significantly increases the computational load of MIM. Building on these advancements, Wu et al. [18] introduced the MotionGRU unit, which can be effectively applied to various RNNs to enhance the ability to model transient changes and motion trends in radar echoes. In addition, Geng et al. [19] applied a generative adversarial network (GAN) architecture, proposing the GAN-rcLSTM that alleviates the issue of blurry echo images. Although the above models have made significant progress in improving extrapolation performance, the inherent recursive characteristics of RNNs in predicting future echo images can lead to the continuous accumulation of errors and the progressive loss of information from distant-past time steps, thereby resulting in high-threshold echo fading and gradual image blurring.

Compared with RNNs, convolutional neural networks (CNNs) are more efficient at capturing spatial features when processing spatiotemporal sequence data, and they generally offer higher computational efficiency. In recent years, CNN-based models have also been used in spatiotemporal prediction tasks. As a case in point, the 3D ConvNet [20] has significantly advanced video prediction tasks. By extending the traditional 2D CNN architecture to process video data directly, it can capture both spatial and temporal information of video data for predicting future events. In contrast, the UNet, proposed by Ronneberger et al. [21], was initially applied in medical image segmentation. However, many meteorological researchers have applied it to radar echo extrapolation [22–24], showing good results. Building upon UNet, the SE-ResUNet by Song et al. [25] incorporates the residual structure of ResNet to increase forecasting accuracy and employs attention mechanisms to learn temporal information. Nevertheless, CNNs struggle to effectively capture long-term temporal dependencies, which limits their effectiveness in predicting complex spatiotemporal data scenarios.

Recently, the Transformer has made a big splash in natural language processing, computer vision, and other fields due to its excellent modeling capability for long-term dependencies and global information acquisition. Given the spatiotemporal nature of meteorological data and the Transformer's proficiency in handling long-sequence data, researchers are continuously exploring the application of the Transformer architecture in weather forecasting. For instance, the Rainformer, designed by Bai et al. [26], utilizes

an improved Transformer structure as a global module to enhance the precision of high-intensity rainfall prediction. Unfortunately, the predictive time of this model is relatively short. Pathak et al. [27] proposed the FourCastNet for global weather forecasting, employing the Transformer architecture based on Fourier neural operators to improve the spatial resolution and accuracy of global precipitation nowcasting. However, this model requires substantial high-resolution data to achieve optimal training performance. Geng et al. [28] introduced the MS-RadarFormer, a model based on Transformer. Its multi-scale design enables this model to better predict the evolution trends of echoes. Unlike previous Transformer-based models, the SFTformer model developed by Xu et al. [29] adopts a hierarchical correlation-decoupling architecture to reduce the interference between spatial and temporal features, thereby improving the prediction accuracy. Overall, Transformer-based models usually require extensive computational resources and data for training, which makes their application in resource-limited environments or on small datasets difficult. In addition, large-scale parameters and complex model structures increase the risk of overfitting.

To address the issues of error accumulation in RNNs, the difficulty of capturing long-term dependencies in CNNs, and the high computational cost associated with Transformer-based models, we propose a non-autoregressive radar echo extrapolation model named MBFE-UNet. This model's overall design adopts the UNet architecture, consisting of two parts: the encoder and the decoder. The core advantage of the MBFE-UNet model stems from its innovative Multi-Branch Feature Extraction Block, which can fully extract spatiotemporal features of radar echoes from three branches and retain more detailed information. In addition, we designed the Temporal Cross Attention Fusion Unit to enable the model to acquire more comprehensive temporal information by fusing low-level and high-level features from the temporal dimension. By integrating these two modules, our model demonstrates superior performance in radar echo extrapolation, effectively mitigating issues such as attenuation of high-threshold echoes and image blurring.

The remainder of this paper is organized as follows: Section 2 introduces the dataset used in the experiments. Section 3 demonstrates our designed model. The implementation details, evaluation metrics, and experimental results are presented in Section 4. Finally, Section 5 provides a summary and discussion.

2. Data

The radar dataset used in this study was obtained after quality control and networking of several S-band weather radars in South China. As shown in Figure 1, the radar data cover the area from 107.5°E to 116.5°E and 18.5°N to 26.5°N. This region falls under the subtropical monsoon climate zone, characterized by abundant rainfall, making it an ideal area for studying radar echo extrapolation. The data cover the period from 2018 to 2022. We used the data from 2018 to 2021 as the training set and the data from 2022 as the validation set to evaluate the model's performance. The radar reflectivity factor values range from 0 to 70, with a horizontal resolution of 0.01° (about 1 km) and a temporal resolution of 6 min. The grid size for single-time data is 800 × 900 pixels.

In terms of data preprocessing, we first manually screened the visualized echo data. On the one hand, we considered data with low echo coverage and reflectivity values below 20 dBZ as clear-sky echoes. We retained only a subset of samples with these characteristics to enhance the model's robustness. On the other hand, we excluded data that had undergone quality control but still exhibited inadequate clutter removal, thereby avoiding negatively impacting the model's training performance. Subsequently, considering the large size of the original images and limited computational resources, we employed linear interpolation to downsample the data to a resolution of 256 × 256 pixels. Finally, a sliding window with a width of 20 frames was used to partition each radar echo sequence. This process resulted in each sub-sequence containing 20 frames of images. The first 10 frames are input for the model, whereas the subsequent 10 frames are the prediction targets. We obtained 14,484 sequences from the training set and 4192 from the validation set. All the

data were normalized before being put into the model to control their values between 0 and 1.

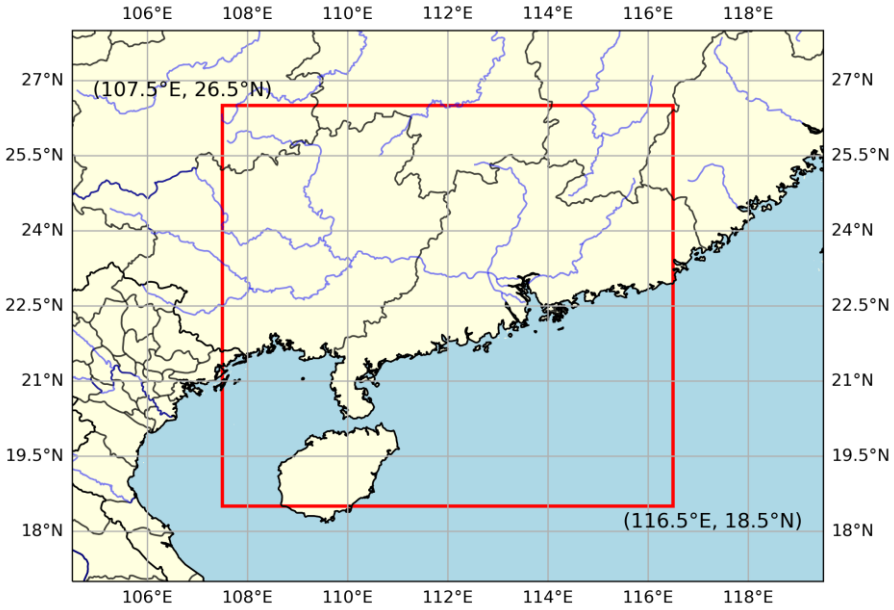

Figure 1. Radar data coverage area (red rectangle).

3. Methods

This section details the MBFE-UNet model. Initially, the overall extrapolation architecture is outlined. Subsequently, the three branches within the Multi-Branch Feature Extraction Blocks are described. Finally, the mechanism by which the Temporal Cross Attention Fusion Unit computes the temporal correlation between encoder and decoder features is elucidated.

3.1. MBFEB-UNet Overall Architecture

The overall architecture of MBFE-UNet is depicted in Figure 2. This model takes 10 consecutive frames (1 h) of radar echo images as input and predicts the subsequent 10 frames (1 h). The key components of the architecture include the encoder, the decoder, and the Temporal Cross Attention Fusion Unit (TCAFU).

The left part of the model is the encoder, which is responsible for extracting spatiotemporal features from radar echo sequences. Initially, the encoder employs the 3D Patch Embedding layer and a 3D convolution layer ($2 \times 4 \times 4$ kernel and stride) to perform preliminary spatiotemporal feature extraction and dimensionality reduction on the radar echo data, mapping the number of input channels to a predefined value C (the default is 32). In the subsequent three stages, Multi-Branch Feature Extraction Blocks (MBFEBs) further extract spatiotemporal features. A Patch Merging layer [30] is applied at the end of the first two stages to losslessly downsample the spatial dimensions of the features (halving both height and width) while doubling the number of channels. As illustrated in Figure 3a, the input features are divided into four segments by the Patch Merging layer, followed by a concatenation along the channel dimension. After normalization, a linear layer is employed to regulate the number of output feature channels. The numbers of MBFEBs in these three stages are [2, 2, 1], with channel counts of [C, 2C, 4C].

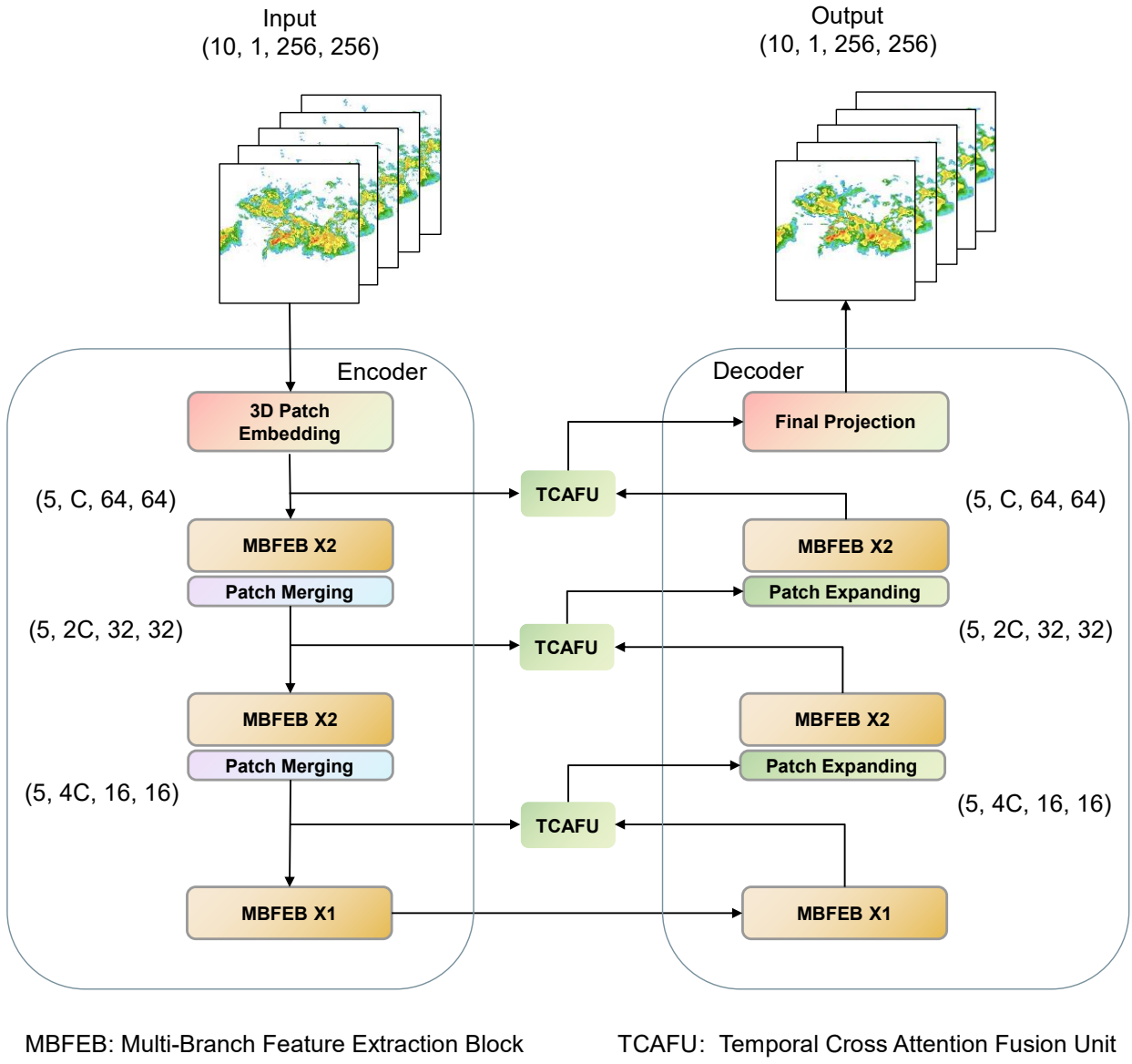

Figure 2. The overall architecture of MBFE-UNet.

The right part of the model is the decoder, which is primarily responsible for progressively restoring the features extracted by the encoder to the original image's spatiotemporal dimensions. The output of the decoder is the sequence of future echo images. Corresponding to the encoder, the decoder is symmetrically constructed based on the MBFEBs. In the first three stages of the decoder, we employ [1, 2, 2] MBFEBs, respectively, with channel counts of [4C, 2C, C]. In contrast, we use the inverse operation of the Patch Merging layer [31], termed the Patch Expanding layer, as the upsampling module within the decoder. As shown in Figure 3b, the Patch Expanding layer first employs a linear layer to double the number of channels of the input features. Then, a rearrange operation is used to reconfigure the features, expanding the spatial resolution of the original input features (2× upsampling) and reducing the channel dimension to half of the original. Finally, the Final Projection layer, a 3D transpose convolution layer ($2 \times 4 \times 4$ kernel and stride), restores the number of feature channels and the spatiotemporal scale to match the original input size, aligning with the model's prediction targets.

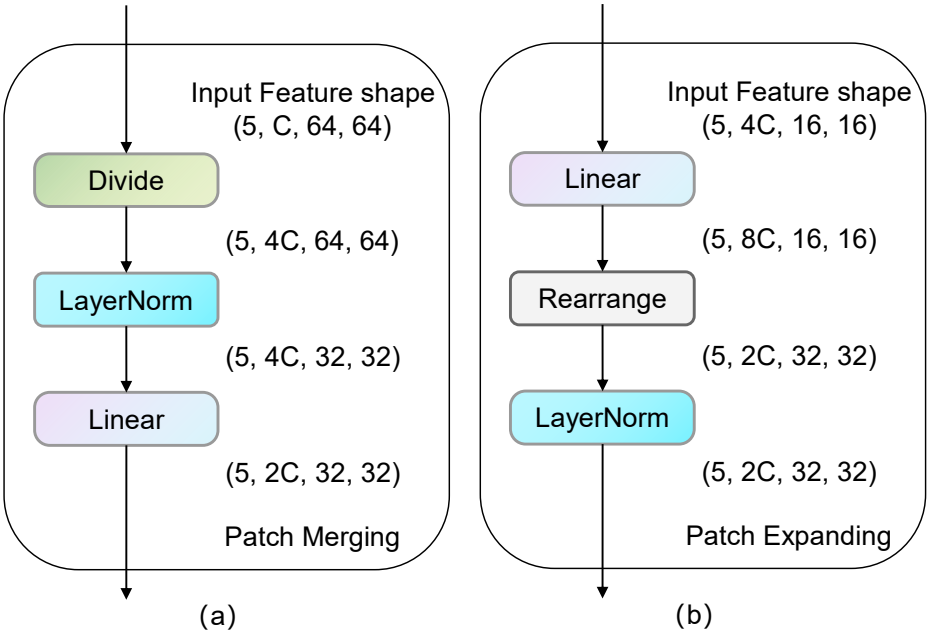

Figure 3. (**a**) presents the Patch Merging operation from the upper left section of Figure 2, while (**b**) illustrates the Patch Expanding operation from the lower right section of Figure 2.

The TCAFU integrates high-level and low-level features and passes the fused features to the subsequent decoding process.

3.2. Multi-Branch Feature Extraction Block

High-quality feature extraction enables the model to capture complex spatiotemporal evolution patterns from radar echo data, thereby enhancing extrapolation performance. Therefore, we designed the MBFEB to synthesize feature information from different perspectives to improve the model's prediction performance. As illustrated in Figure 4, the MBFEB has three branches: a local branch, a sensitive branch, and a global branch.

Accurately capturing the variation trends of radar echoes within short time frames and small spatial regions is crucial for predicting localized extreme weather events. To better maintain the local fidelity of the extrapolation results, we designed local branches within the MBFEB to endow the model with local perception capabilities. This branch includes a 3D convolutional layer ($3 \times 3 \times 3$ kernel), batch normalization (BN), ReLU activation function, and a multi-layer perceptron (MLP). The 3D convolution operation extracts fine-grained features from both spatial and temporal dimensions of radar echoes, thereby capturing the short-term variations in the intensity and morphology of echoes within localized regions. We incorporated batch normalization (BN) before the ReLU activation function to normalize each batch, accelerating the model's convergence and stabilizing the training process. Finally, the MLP is employed to regulate the number of output channels for the branch.

In order to enhance the model's focus on strong echo features, we designed the sensitive branch to extract the most prominent features within localized spatiotemporal regions. This branch consists of a MaxPooling 3D layer ($3 \times 3 \times 3$ kernel), an UpSample layer, and an MLP layer. Initially, the Maxpooling 3D operation retains the maximum value within the pooling window, enabling the model to focus on the most sensitive features in radar echo data. Following this, to ensure that the output shape is consistent with the other two branches, an UpSample layer is used to upsample the spatiotemporal scale of the feature map back to its input shape of this branch. Finally, the MLP further extracts high-order representations of the sensitive features.

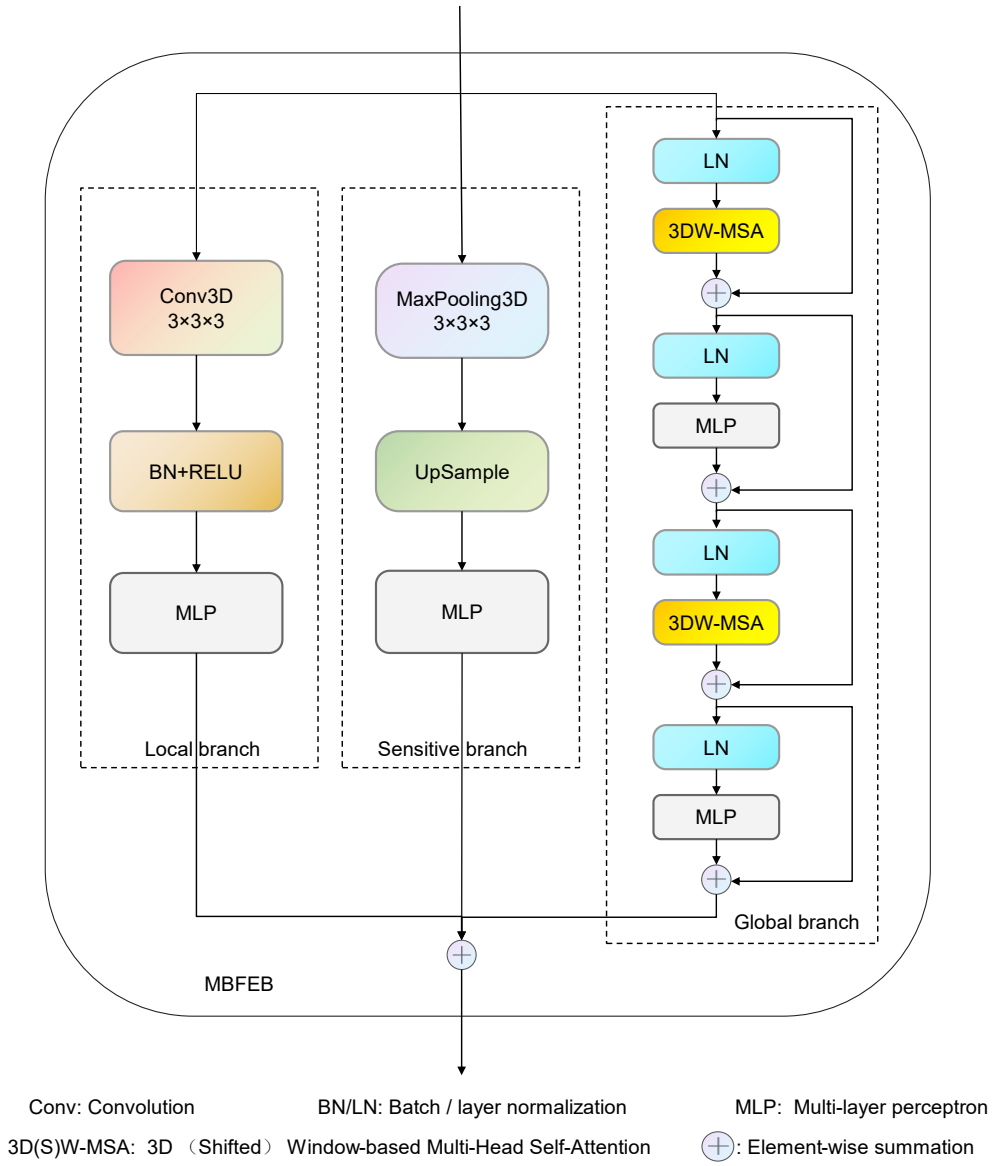

Figure 4. The structure of the Multi-Branch Feature Extraction Block.

To better capture the evolution trends of large-scale echo, the MBFEB integrates a global branch to model the global spatiotemporal correlations of radar echoes. Due to the temporal dimension in radar echo data, the traditional self-attention [32] would incur substantial computational costs. Therefore, we employ the 3D Window-based Multi-Head Self-Attention (3D W-MSA) and 3D Shifted Window-based Multi-Head Self-Attention (3D SW-MSA) from the Video Swin Transformer Block [33] as feasible alternatives. As illustrated in the left part of Figure 5, the 3D W-MSA divides the input feature maps into multiple non-overlapping 3D windows and performs multi-head attention computation within each window. However, this approach is inherently limited by the inability to facilitate information exchange between windows. To establish connections between independent windows, 3D SW-MSA employs a shifted window partitioning approach to divide the feature maps again. In the newly partitioned windows, attention computation is performed again, thereby expanding the range of attention. Overall, the combined application of 3D W-MSA and 3D SW-MSA facilitates global feature extraction from radar echo data while reducing computational overhead.

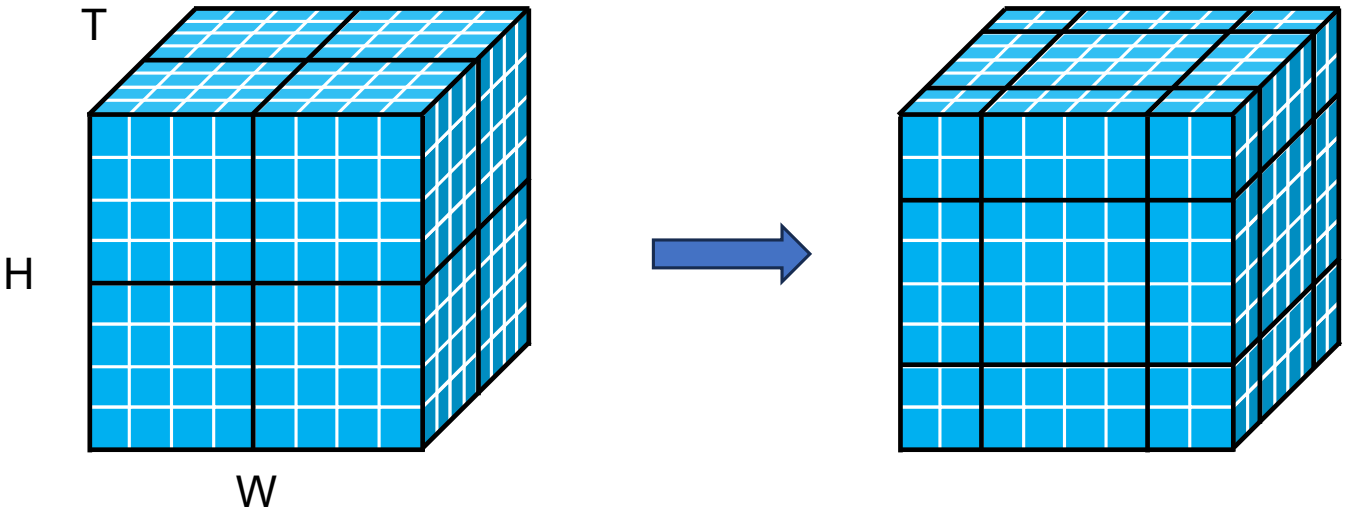

Figure 5. An illustrated example of 3D Window-based Multi-Head Self-Attention (3D W-MSA) and 3D Shifted Window-based Multi-Head Self-Attention (3D SW-MSA). T represents the temporal dimension, while H and W represent the spatial dimension.

Within localized spatiotemporal regions, the local branch and sensitive branch, respectively, learn fine-grained features of strong echoes and capture coarse-resolution information. By utilizing attention mechanisms, the global branch captures the evolutionary features of strong echoes over larger spatiotemporal scales. These features play a guiding role in improving the accuracy of strong echo predictions during extrapolation, effectively mitigating the attenuation of high-threshold echoes.

3.3. Temporal Cross Attention Fusion Unit

To preserve more extrapolation details and enhance the model's ability to learn the temporal information of radar echoes, we designed the Temporal Attention Fusion Unit (TCAFU), as shown in Figure 6. As the network depth increases, the detail information retained in the low-level features gradually decreases, while the high-level features contain rich details of the echo images. Unlike the straightforward channel concatenation strategy employed in traditional UNet architectures, the TCAFU utilizes a cross-attention mechanism to capture the intricate relationships between encoder and decoder features.

Before being fed into the TCAFU, the two features are reshaped from (b, t, h, w, c) to (b $\times$ h $\times$ w, t, c), which involves merging the spatial dimensions and preserving the temporal dimension. This transformation helps the model learn the temporal dependencies of radar echoes at the same spatial location across different time steps. The subsequent computational process is as follows: First, the features from the encoder are linearly transformed to generate a query matrix (Q), and decoder features are similarly transformed to produce a key matrix (K) and a value matrix (V). Next, Q is multiplied by K to compute the attention weights, which measure the correlation between the encoder and decoder features in the temporal dimension. These weights are then scaled and normalized. Finally, the attention weights are multiplied by V to complete the feature fusion.

The above process fuses the two types of features along the temporal dimension and improves the issue of detail loss in the decoder features. Fine-grained information from the encoder features is continuously incorporated into the radar echo image reconstruction, which helps the model enhance the clarity of the extrapolated images.

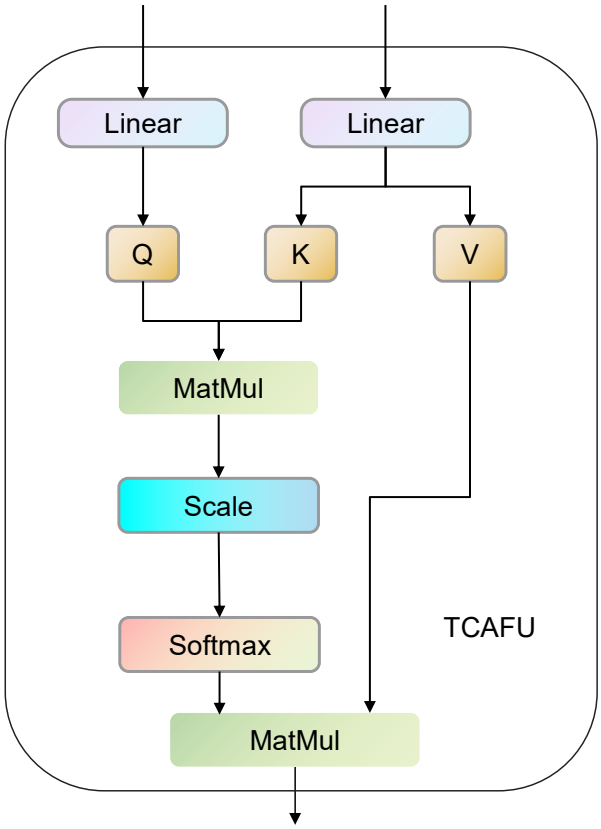

Figure 6. The structure of Temporal Cross Attention Fusion Unit.

4. Experiments and Analysis

4.1. Implementation Details

In this study, all experiments were implemented by the PyTorch deep learning framework and conducted on an NVIDIA RTX 4090 GPU (24 GB) manufactured by Colorful Group in Shenzhen, China. All models were trained for 50 epochs to ensure fairness in comparison, which was sufficient for all models to converge. The training batch size was set to 4, and the initial learning rate was set to 10^{-3}. Each model was optimized with Adam optimizer. We chose a combined loss function of Mean Squared Error (MSE) and Mean Absolute Error (MAE) because the exclusive use of MSE may result in excessively small loss values that negatively impact the training effect, while MAE may lead to the loss of essential data feature details [34]. The formula for the loss function is as follows:

$$\mathrm{MSE} = \frac{1}{\mathrm{T} \times \mathrm{H} \times \mathrm{W}} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(\mathrm{Y}_{t,h,w} - \widetilde{\mathrm{Y}}_{t,h,w} \right)^2 \tag{1}$$

$$\mathrm{MAE} = \frac{1}{\mathrm{T} \times \mathrm{H} \times \mathrm{W}} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{w=1}^{W} \left| \mathrm{Y}_{t,h,w} - \widetilde{\mathrm{Y}}_{t,h,w} \right| \tag{2}$$

$$\mathrm{Loss} = \mathrm{MSE} + \mathrm{MAE} \tag{3}$$

where $\mathrm{Y}_{t,h,w}$ denotes the actual radar echo value of the target image sequences with pixel coordinates (h,w) at timestamp t, and $\widetilde{\mathrm{Y}}_{t,h,w}$ is the corresponding predicted value. T is the total length of the predicted sequence, while H and W are the height and width of the radar image, respectively.

4.2. Evaluation Metrics

In order to comprehensively evaluate the performance of the models in the radar echo extrapolation task, we used three standard meteorological evaluation metrics: crit-

ical success index (CSI) [35], probability of detection (POD) [36], and Heidke skill score (HSS) [37], along with the structural similarity index measure (SSIM) [38]. These assessment metrics objectively measure the predictive power and accuracy of the model and reflect the consistency between predicted results and actual observations.

To compute these metrics, we converted the predicted and ground truth images into binary matrices by a given threshold τ, representing a corresponding echo intensity level. In detail, values in the predicted or actual images above τ were set to 1, otherwise to 0. Based on binary matrices, we can construct a confusion matrix. In binary classification tasks, the confusion matrix is a commonly used tool for evaluating model performance, presenting the correspondence between the model's predictions and actual labels in a tabular format. As shown in Table 1, the confusion matrix includes the counts of true positives (prediction = 1, truth = 1, denoted as TP), false positives (prediction = 1, truth = 0, denoted as FP), true negatives (prediction = 0, truth = 0, denoted as TN), and false negatives (prediction = 0, truth = 1, denoted as FN). These counts enable the calculation of metrics such as CSI, POD, and HSS. In our experiments, we selected thresholds of 10 dBZ, 20 dBZ, and 30 dBZ.

Table 1. Confusion matrix.

	Prediction: 1	**Prediction: 0**
Truth: 1	TP (true positive)	FN (false negative)
Truth: 0	FP (false positive)	TN (true negative)

The CSI, also known as the threat score (TS), is an essential measure of the accuracy of predicted echo strength. The CSI value is calculated as follows:

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{4}$$

The POD reflects the proportion of correct events predicted by the model to all actual correct events, which is calculated as follows:

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

HSS measures the predictive efficacy of the model by comparing the difference between the actual and stochastic prediction. The formula for calculating the HSS is as follows:

$$\mathrm{HSS} = \frac{2 \times (\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP})}{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{FN} + \mathrm{TN}) + (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{FP} + \mathrm{TN})} \tag{6}$$

The values of CSI and POD range from 0 to 1, and the HSS value varies between -1 and 1. The CSI, POD, and HSS can intuitively reflect the model's performance, with higher values indicating better predictive performance.

The SSIM is a widely used metric for assessing image quality, especially in image reconstruction and enhancement tasks. The index ranges from -1 to 1, where a value of 1 indicates perfect structural similarity. The definition of SSIM is as follows:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{7}$$

where x and y represent the two images being compared, μ_x and μ_y denote the mean values of x and y, while σ_x^2 and σ_y^2 indicate the variances of the two images. The term σ_{xy} represents the covariance between x and y. To stabilize the division in the formula, constants c_1 and c_2 are introduced with c_1 defined as $c_2 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$. Here, L is the dynamic range of pixel values. $k_1 = 0.01$ and $k_2 = 0.03$.

4.3. Comparative Study

To demonstrate the superiority of our model, we conducted a performance comparison of the MBFE-UNet model with a series of advanced models, including ConvLSTM, PredRNN, MotionRNN, MIM, 3D-UNet, and Rainformer. The input and output both consist of 10 frames of radar echo images. To ensure the fairness of the experiment, all models used the same optimizer and learning rate and were trained for the same number of epochs.

To quantitatively evaluate the extrapolation performance of the model, we computed the average CSI, POD, HSS, SSIM, and MSE values across all time steps for all models on the validation set. Tables 2–4 present the CSI, POD, and HSS scores at thresholds of 10 dBZ, 20 dBZ, and 30 dBZ, respectively, and Table 5 summarizes the results of SSIM and MSE. The best and second-best scores are marked in bold and underlined, respectively. From the results of the tables, the MBFE-UNet model achieved the best scores in CSI, POD, and HSS at most thresholds. Additionally, as the threshold increases, the superiority of MBFE-UNet becomes increasingly evident. This indicates that our model effectively extracts spatiotemporal features from radar echoes, thereby enhancing the accuracy of echo extrapolation, particularly for strong echoes. Compared to Rainformer, our model exhibits an average improvement of 4.8% on CSI, an average improvement of 5.5% on POD, and an average improvement of 3.8% on HSS. Among the RNNs, ConvLSTM scored the lowest across all metrics at all thresholds. PredRNN and MotionRNN showed similar performance, slightly outperforming ConvLSTM. MIM has a higher prediction accuracy than the first three models, especially at a threshold of 30 dBZ, which indicates that it is more capable of predicting strong echoes. The overall scores of 3D-UNet and Rainformer were superior to the RNN-based models. As shown in Table 5, MBFE-UNet achieves the highest SSIM score, indicating that the extrapolated images generated by our model exhibit the highest structural similarity to the ground truth images.

Table 2. The average CSI scores of 10 frames for MBFE-UNet and other models at the thresholds of 10 dBZ, 20 dBZ, and 30 dBZ.

Model	10 dBZ	20 dBZ	30 dBZ	Avg
ConvLSTM	0.6577	0.5795	0.3532	0.5301
PredRNN	0.6642	0.5859	0.3656	0.5386
MotionRNN	0.6646	0.5911	0.3668	0.5408
MIM	0.6665	0.5949	0.3782	0.5465
3D-UNet	<u>0.6936</u>	<u>0.6348</u>	0.4008	<u>0.5764</u>
Rainformer	0.6854	0.6218	<u>0.4073</u>	0.5715
MBFE-UNet	**0.7023**	**0.6450**	**0.4490**	**0.5988**

The best and second-best scores are marked in bold and underlined, respectively. The 'Avg' represents the average scores of 10 dBZ, 20 dBZ, and 30 dBZ.

Table 3. The average POD scores of 10 frames for MBFE-UNet and other models at the thresholds of 10 dBZ, 20 dBZ, and 30 dBZ.

Model	10 dBZ	20 dBZ	30 dBZ	Avg
ConvLSTM	0.7258	0.6668	0.4112	0.6013
PredRNN	0.7325	0.6751	0.4274	0.6117
MotionRNN	0.7318	0.6791	0.4259	0.6122
MIM	0.7304	0.6838	0.4432	0.6191
3D-UNet	**0.7792**	0.6899	0.4335	0.6342
Rainformer	0.7469	<u>0.7008</u>	<u>0.4648</u>	<u>0.6375</u>
MBFE-UNet	<u>0.7620</u>	**0.7277**	**0.5275**	**0.6724**

The best and second-best scores are marked in bold and underlined, respectively. The 'Avg' represents the average scores of 10 dBZ, 20 dBZ, and 30 dBZ.

Table 4. The average HSS scores of 10 frames for MBFE-UNet and other models at the thresholds of 10 dBZ, 20 dBZ, and 30 dBZ.

Model	10 dBZ	20 dBZ	30 dBZ	Avg
ConvLSTM	0.7561	0.7040	0.4890	0.6497
PredRNN	0.7614	0.7092	0.5037	0.6581
MotionRNN	0.7619	0.7138	0.5045	0.6601
MIM	0.7640	0.7177	0.5193	0.6670
3D-UNet	**0.8019**	<u>0.7511</u>	0.5380	<u>0.6970</u>
Rainformer	0.7789	0.7398	<u>0.5446</u>	0.6878
MBFE-UNet	<u>0.7955</u>	**0.7591**	**0.5868**	**0.7138**

The best and second-best scores are marked in bold and underlined, respectively. The 'Avg' represents the average scores of 10 dBZ, 20 dBZ, and 30 dBZ.

Table 5. The average SSIM and MSE values of 10 frames for MBFE-UNet and other models.

Model	SSIM	MSE×100
ConvLSTM	0.8698	0.4157
PredRNN	0.8726	0.4082
MotionRNN	0.8735	0.4018
MIM	0.8751	0.3965
3D-UNet	<u>0.8884</u>	<u>0.3403</u>
Rainformer	0.8832	0.3637
MBFE-UNet	**0.8898**	**0.3325**

The best and second-best scores are marked in bold and underlined, respectively.

To visually compare the extrapolation performance of strong echoes at different time steps, we present the CSI, POD, and HSS variation curves for the MBFE-UNet and advanced models at the threshold of 30 dBZ in Figure 7. The results show that the predictive performance of all models keeps deteriorating as the extrapolation time step increases. This suggests that accurately predicting echoes over longer time steps is more challenging. As illustrated in Figure 7, our model achieved the highest CSI, POD, and HSS scores across all prediction time steps. This indicates that our model provides more accurate predictions for regions with strong echoes. The performances of ConvLSTM, PredRNN, and MotionRNN are quite similar. Although MIM's performance is slightly lower than that of PredRNN and MotionRNN at the initial time step, it gradually demonstrates superiority as the time step increases. This could be attributed to MIM's ability to capture the stationary and non-stationary information in the echo data. The variation curves of CSI and HSS indicate that 3D-UNet and Rainformer also exhibited similar performance; however, Rainformer achieved higher POD scores. Compared with the RNN-based models, MBFE-UNet is based on the non-autoregressive architecture, which means that this model is unaffected by the error accumulation problem during the extrapolation process. Consequently, our model maintains good performance even in the later stages of prediction.

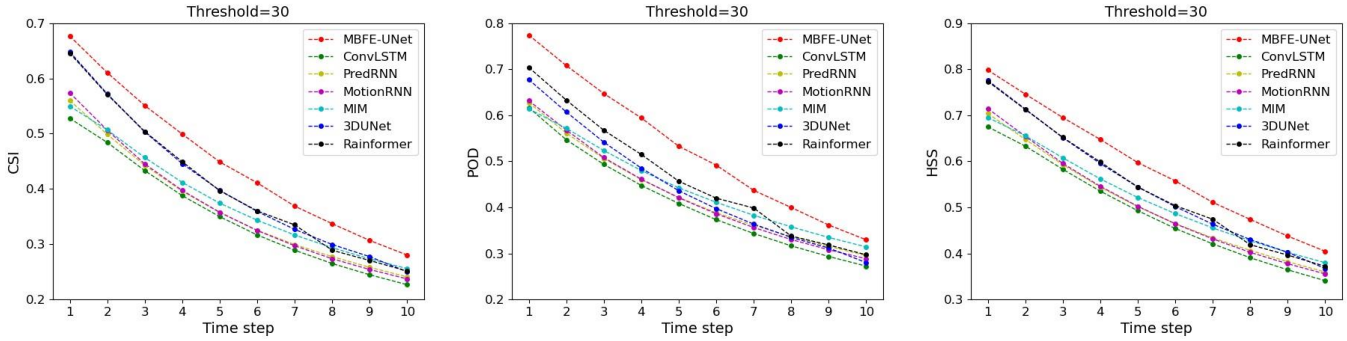

Figure 7. CSI, POD, and HSS curves at the 30 dBZ along different extrapolation time steps.

To qualitatively assess the predictive performance of the model, we visualized two examples in Figures 8 and 9. In the first five extrapolation time steps, the MBFE-UNet demonstrated a significant advantage. Its predicted images exhibited the highest clarity, and the echo intensities closely matched the ground truth, particularly for strong echoes. Figure 8 illustrates a squall line weather event moving northeastward, characterized by a relatively concentrated distribution. As the time steps progressed, MBFE-UNet demonstrated superior performance compared with other models in maintaining the intensity of red echoes above 45 dBZ in the central region of the images. The ConvLSTM exhibits limitations in predicting strong echoes at T = 11 and undergoes severe decay starting from the fifth time step. PredRNN, MotionRNN, and MIM perform similarly and better than ConvLSTM. However, at T = 19, the spatial morphology of the yellow echoes in the upper left corner deviates significantly from the ground truth, and there is severe attenuation of the dark yellow echoes above 35 dBZ in the lower right corner. Compared with RNN-based models, 3D-UNet and Rainformer can accurately predict the red echo regions above 45 dBZ. However, starting from T = 17, 3D-UNet shows a significant deterioration in its predictions for echoes within the range of 35–40 dBZ. Figure 9 illustrates another severe convective weather event. As observed from the ground truth in the first row, the yellow echo region on the left part of the image continually expands. The MBFE-UNet model accurately captures this evolution trend and maintains the spatial morphology of the echo region. In contrast, predictions from other models exhibit a weakening trend and, with increasing extrapolation steps, lose the crescent-shaped structure of the region. In addition, the MBFE-UNet maintains more intensity information for the red strong echo region above 45 dBZ at T = 19, whereas the other models can hardly predict echoes of this intensity.

Visual analysis reveals three notable advantages of the MBFE-UNet model in the radar echo extrapolation task. Firstly, MBFE-UNet demonstrates superior accuracy in predicting high-intensity echo regions. Secondly, the model excels in preserving the spatial morphology of echoes in the later stages of prediction. Finally, our model is capable of better predicting the evolution trends of large-scale radar echoes. The advantages above highlight the model's robust feature extraction capabilities and its effectiveness in capturing the spatiotemporal evolution patterns of radar echoes.

4.4. Ablation Study

In this section, we validated the effectiveness of each branch within the MBFEB and the TCAFU through a series of ablation experiments. For the experiments on the MBFEB, we sequentially removed one branch at a time while retaining the other two branches intact. In the case of removal of the TCAFU, we concatenated the features transmitted via the skip connections with the features of the decoder along the channel dimension. The scoring results of the experiments are detailed in Tables 6–9. The highest and second-highest scores are marked in bold and underlined, respectively. The data in the tables show that the full MBFE-UNet achieved the highest scores in most metrics compared with the other ablation models. Additionally, an extrapolation example of the original model and the ablation models is presented in Figure 10.

We chose UNet as the baseline model. Compared to the baseline, at a threshold of 30 dBZ, the MBFE-UNet achieves improvements of 10%, 12.5%, and 9.7% in CSI, POD, and HSS, respectively. This indicates that our model demonstrates superior predictive performance for strong echoes. The visualization case also reflects this point. In the later stages of prediction, UNet's ability to predict echoes above 35 dBZ is inferior to that of MBFE-UNet.

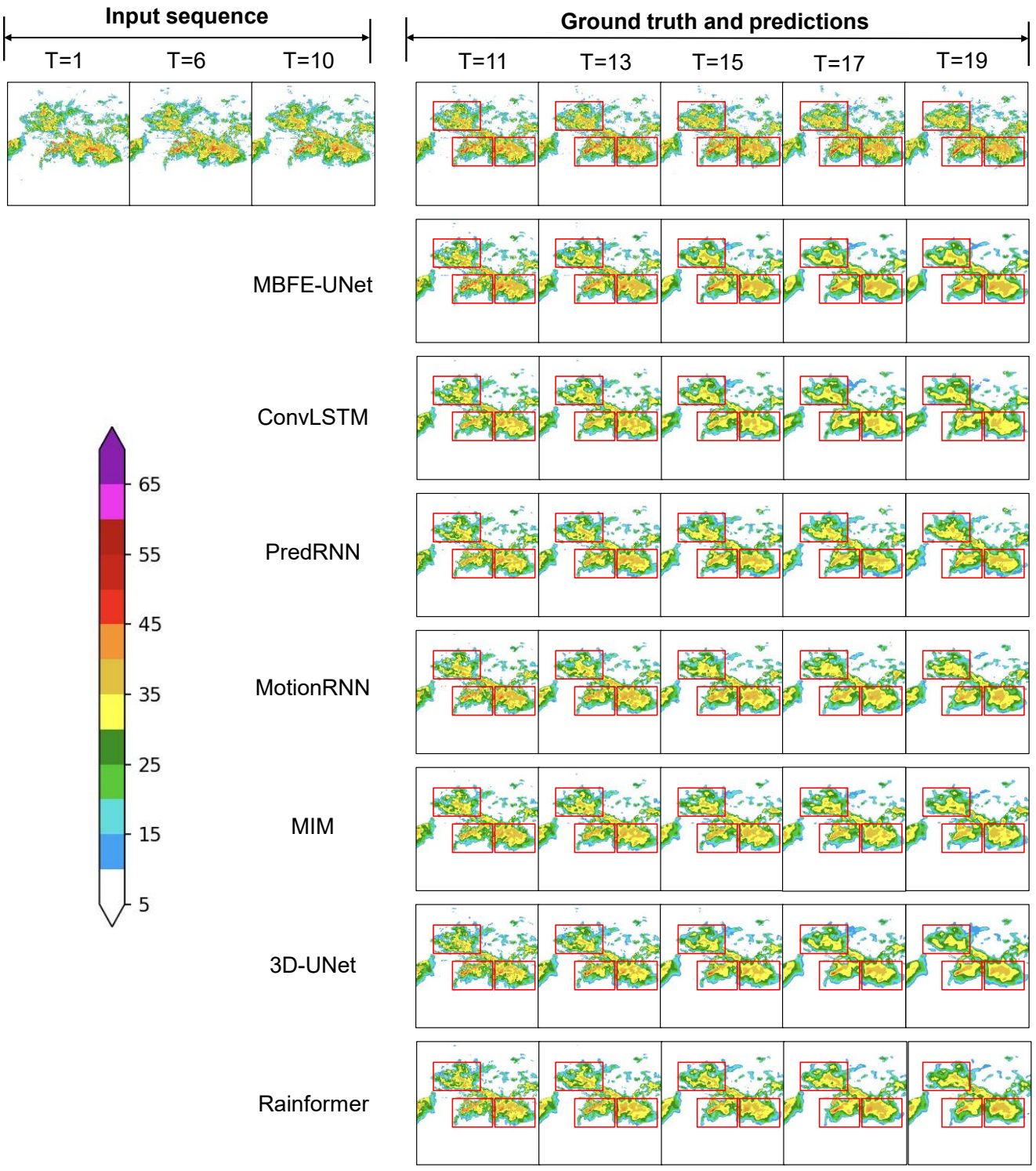

Figure 8. Extrapolation example 1 of all models. The first row is the ground truth. Rows two to eight, respectively, show the results of MBFE-UNet, ConvLSTM, PredRNN, MotionRNN, MIM, 3D-UNet, and Rainformer. The red rectangles highlight the key comparison area.

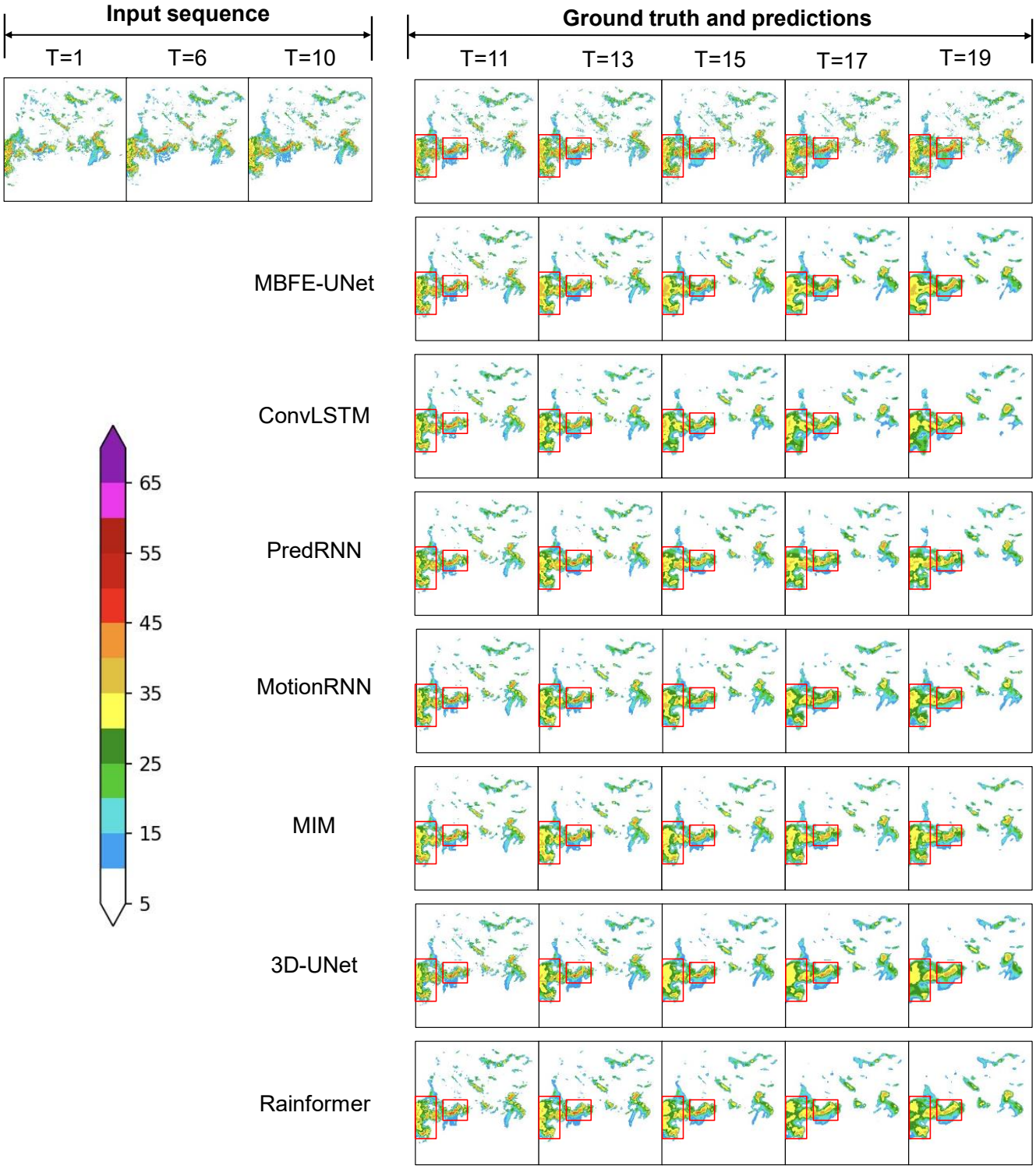

Figure 9. Extrapolation example 2 of all models. The first row is the ground truth. Rows two to eight, respectively, show the results of MBFE-UNet, ConvLSTM, PredRNN, MotionRNN, MIM, 3D-UNet, and Rainformer. The red rectangles highlight the key comparison area.

Table 6. The average CSI scores of 10 frames for MBFE-UNet and ablated models at the thresholds of 10 dBZ, 20 dBZ, and 30 dBZ.

Model	10 dBZ	20 dBZ	30 dBZ	Avg
UNet	0.6984	0.6317	0.4050	0.5784
Only B-L	0.6983	0.6291	0.4130	0.5801
Only B-S	0.4424	0.2853	0.0958	0.2745
Only B-G	0.6944	0.6326	0.4235	0.5835
W/O B-L	0.6862	0.6247	0.4330	0.5813
W/O B-S	0.7002	0.6429	0.4457	0,5962
W/O B-G	0.7004	0.6308	0.4137	0.5816
W/O TCAFU	**0.7057**	0.6434	0.4428	0.5973
MBFE-UNet	0.7023	**0.6450**	**0.4490**	**0.5988**

The best and second-best scores are marked in bold and underlined, respectively. The 'Avg' represents the average scores of 10 dBZ, 20 dBZ, and 30 dBZ.

Table 7. The average POD scores of 10 frames for MBFE-UNet and ablated models at the thresholds of 10 dBZ, 20 dBZ, and 30 dBZ.

Model	10 dBZ	20 dBZ	30 dBZ	Avg
UNet	0.7645	0.7184	0.4687	0.6505
Only B-L	0.7610	0.7028	0.4669	0.6436
Only B-S	0.5060	0.3248	0.1126	0.3145
Only B-G	0.7586	0.7254	0.4986	0.6609
W/O B-L	0.7521	0.7202	0.5258	0.6660
W/O B-S	0.7601	**0.7315**	0.5262	**0.6726**
W/O B-G	0.7644	0.7061	0.4706	0.6470
W/O TCAFU	**0.7718**	0.7288	0.5164	0.6723
MBFE-UNet	0.7620	0.7277	**0.5275**	0.6724

The best and second-best scores are marked in bold and underlined, respectively. The 'Avg' represents the average scores of 10 dBZ, 20 dBZ, and 30 dBZ.

Table 8. The average HSS scores of 10 frames for MBFE-UNet and ablated models at the thresholds of 10 dBZ, 20 dBZ, and 30 dBZ.

Model	10 dBZ	20 dBZ	30 dBZ	Avg
UNet	0.7893	0.7469	0.5346	0.6903
Only B-L	0.7901	0.7461	0.5494	0.6952
Only B-S	0.5500	0.3939	0.1502	0.3647
Only B-G	0.7868	0.7486	0.5590	0.6981
W/O B-L	0.7801	0.7422	0.5720	0.6981
W/O B-S	0.7915	0.7572	0.5832	0.7106
W/O B-G	0.7915	0.7469	0.5476	0.6953
W/O TCAFU	**0.7957**	0.7577	0.5808	0.7114
MBFE-UNet	0.7955	**0.7591**	**0.5868**	**0.7138**

The best and second-best scores are marked in bold and underlined, respectively. The 'Avg' represents the average scores of 10 dBZ, 20 dBZ, and 30 dBZ.

Table 9. The average SSIM and MSE values of 10 frames for MBFE-UNet and ablated models.

Model	SSIM	MSE×100
UNet	0.8864	0.3478
Only B-L	0.8871	0.3412
Only B-S	0.7995	0.8050
Only B-G	0.8856	0.3493
W/O B-L	0.8815	0.3643
W/O B-S	0.8883	0.3386
W/O B-G	0.8877	0.3388
W/O TCAFU	0.8897	0.3357
MBFE-UNet	**0.8898**	**0.3325**

The best and second-best scores are marked in bold and underlined, respectively.

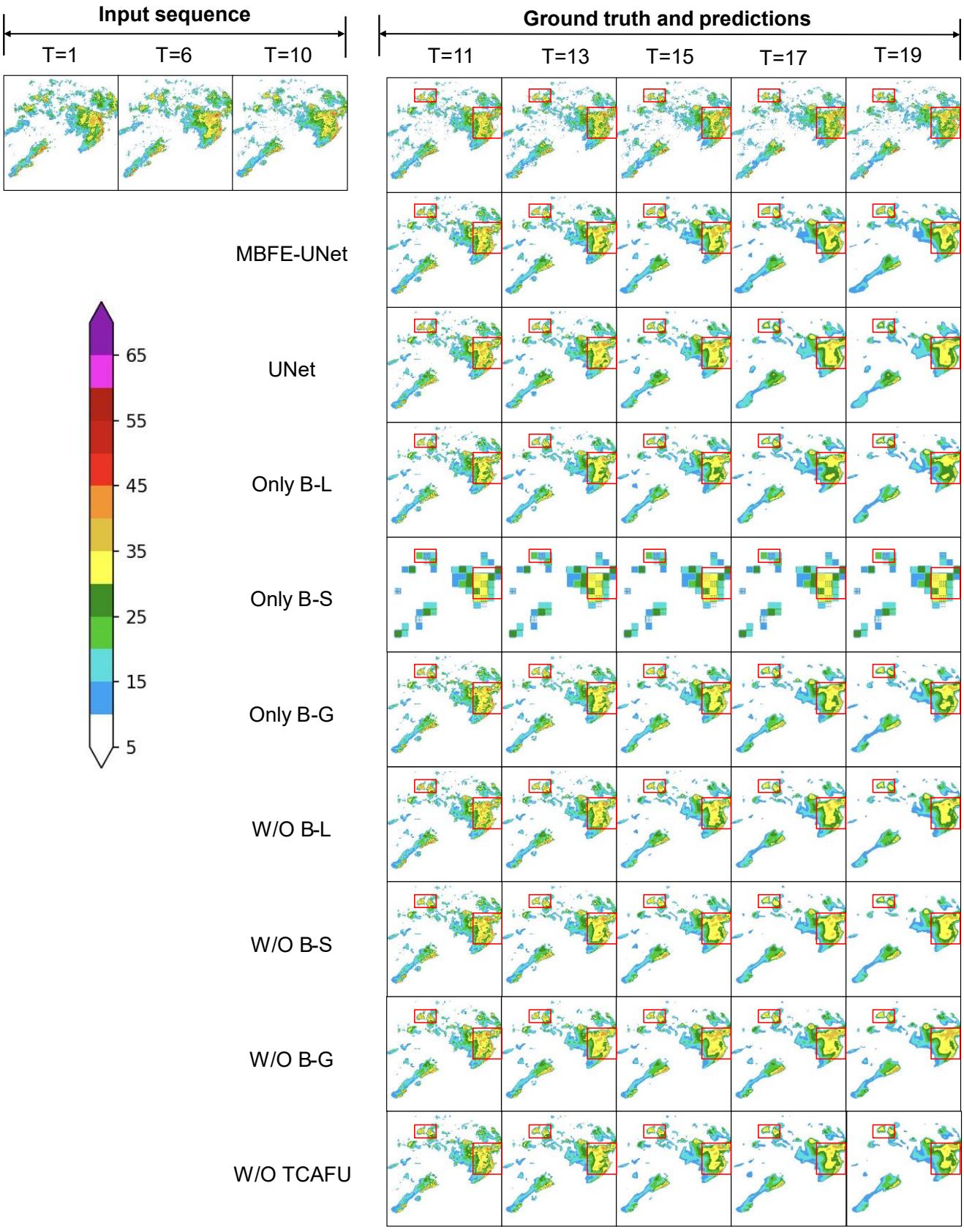

Figure 10. Extrapolation example of MBFE-UNet and ablated models. The first row is the ground truth. Rows two to ten, respectively, show the results of MBFE-UNet, UNet, Only B-S, Only B-L, Only B-G, W/O B-L, W/O B-S, W/O B-G, and W/O TCAFU. The red rectangles highlight the key comparison area.

4.4.1. Multi-Branch Feature Extraction Block

To further explore the impact of local, sensitive, and global features extracted by the MBFEB on predictive performance, we trained six variants of the MBFE-UNet: one with only local branch (denoted as Only B-L), one with only sensitive branch (denoted as Only B-S), and one with only global branch (denoted as Only B-G). one without local branch (denoted as W/O B-L), one without sensitive branch (denoted as W/O B-S), and one without global branch (denoted as W/O B-G).

The experimental results in Tables 6–9 indicate that the complete MBFE-UNet outperforms all scores of the three variants that utilize individual branches. Among these, the performance of Only B-S is the poorest, and it can only predict very coarse-resolution images. The phenomenon occurs because the single sensitive branch can only extract features through the MaxPooling layers at various stages of the network, leading to significant information loss. Compared to MBFE-UNet, the problem of strong echo attenuation is significantly more severe in Only B-L and Only B-G. From the extrapolation results in Figure 10, it can be observed that they are nearly unable to predict areas with echoes exceeding 35 dBZ from T = 17.

As shown in Tables 6–8, MBFE-UNet outperforms W/O B-L under all thresholds. The visualization results in Figure 10 indicate that W/O B-L loses more detailed information in multiple local echo regions from T = 15, which suggests that the local branch is beneficial for extracting the local spatiotemporal features of radar echoes. Although the POD slightly increased after removing the sensitive branch, it resulted in worse CSI and HSS scores. Furthermore, as illustrated in Figure 10, W/O B-S experiences severe echo attenuation. From T = 17, W/O B-S encounters difficulties in predicting the strong echo regions above 40 dBZ in the upper left section of the images. This suggests that the sensitive branch is able to extract more strong echo features. The CSI, POD, and HSS scores of W/O B-G at the threshold of 30 significantly declined, which indicates a deterioration in the model's ability to predict strong echoes. Moreover, in the absence of the global branch, the model's capability to capture the overall trend of echo variations is notably inferior to that of the original model. In the ground truth shown in the first row of Figure 10, the yellow echo region on the right part of the image exhibits a progressive attenuation trend. However, predictions from W/O B-G show a continuously increasing area of this echo region.

In summary, all three branches play crucial roles in extracting the spatiotemporal features of radar echoes and significantly enhance the accuracy of echo extrapolation.

4.4.2. Temporal Cross Attention Fusion Unit

To demonstrate the validity of the TCAFU, we compared the MBFE-UNet model without the TCAFU (denoted as W/O TCAFU) with the complete MBFE-UNet model. Tables 6–9 show that MBFE-UNet outperformed W/O TCAFU across multiple metrics. The difference between the two models was particularly pronounced at the 30 dBZ threshold. In addition, as can be seen from the visualization results in Figure 10, the W/O TCAFU model's prediction of echoes in the 35–40 dBZ range significantly decreased from T = 15, and it could hardly predict echoes of this intensity at T = 19. In contrast, the complete model maintained the echo intensity much better. Therefore, the TCAFU enables the model to extract more temporal features from different network layers, effectively improving issues of strong echo attenuation and better retaining more image details.

5. Conclusions

In recent years, the frequency of severe convective weather events such as intense precipitation, thunderstorms, and hail has increased, posing significant threats to public safety and property. Radar echo extrapolation is a weather forecasting technique that provides timely predictions of extreme weather phenomena. This enables forecasters to identify potential hazard areas in advance and formulate effective response strategies, thereby mitigating the risks associated with such events. However, traditional extrapolation methods often struggle to accurately capture the complex dynamics of severe convective weather,

leading to reduced precision in predicting the intensity and movement of radar echoes. Deep learning extrapolation methods can effectively mine the spatiotemporal information within echo data, enhancing extrapolation performance. Nevertheless, challenges such as attenuation of high-threshold echoes and image blurring persist.

In this study, we propose a novel UNet-based radar echo extrapolation model named MBFE-UNet to mitigate the above problems. We designed the MBFEB to comprehensively extract spatiotemporal features from radar echo data through three distinct branches, enabling the model to better capture the spatiotemporal evolution patterns of radar echoes. Structurally, we introduced the TCAFU to integrate features from the encoder and decoder and enhance the model's capability to exploit temporal information. By integrating the two modules, MBFE-UNet demonstrates superior performance in the task of radar echo extrapolation. Through analyzing the experimental results and visualization cases, the following conclusions are drawn:

1. The Multi-Branch Feature Extraction Block adopts multi-branch architecture to extract spatiotemporal features of radar echoes, effectively improving the prediction accuracy of strong echo regions and large-scale echo evolution trends.
2. The Temporal Cross Attention Fusion Unit can extract more temporal information from the features of different network layers, enabling the model to retain more image details during the extrapolation process.
3. Experimental results demonstrate that the MBFE-UNet model significantly outperforms other advanced methods in radar echo extrapolation. In particular, the model exhibits superior predictive performance for strong echoes, reflecting its enhanced focus on these intensity levels.

Radar echoes are closely linked to convective systems, and the extrapolated echo images can be used to predict the evolution of these systems. Our MBFE-UNet model significantly enhances the accuracy of strong echo predictions and improves image clarity, enabling forecasters to accurately and promptly anticipate severe convective weather such as heavy rainfall, hail, and thunderstorms. This allows for formulating preventive measures against these hazardous events, helping to reduce casualties and economic losses. Thus, the MBFE-UNet model presents significant potential applications in meteorology and emergency preparedness.

Although our MBFE-UNet model performs excellently in the radar echo extrapolation task, the model still has some limitations. First, long-time echo extrapolation still has room for improvement, which remains a significant challenge in weather forecasting. Second, the model mainly relies on single-source radar data, which limits its ability to predict more complex weather processes. In future work, we plan to enhance extrapolation performance by refining the model's architecture and incorporating multi-source data. Specifically, we will explore the adoption of new structures, such as diffusion models, to generate clearer future radar images during long-term extrapolation. Additionally, we will integrate satellite data and ground observation station data to provide a more comprehensive understanding of weather conditions and supplement constraint information.

Author Contributions: Conceptualization, H.Z.; methodology, H.Z.; software, H.Z.; validation, H.Z.; formal analysis, H.Z.; investigation, H.Z., Z.S., F.W., L.G. and K.M.; resources, H.G.; data curation, Z.S.; writing—original draft, H.Z.; writing—review and editing, H.G.; visualization, H.Z.; supervision, H.G.; project administration, H.G.; funding acquisition, H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 42375145; The Open Grants of China Meteorological Administration Radar Meteorology Key Laboratory, grant number 2023LRM-A02; and The Major Science and Technology Special Project Funding of Jiangxi Province, grant number 20223BBG71019.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the confidentiality policy.

Acknowledgments: The authors wish to express their sincere gratitude to the anonymous reviewers for their professional and insightful feedback on this manuscript. Their constructive comments and suggestions have significantly contributed to the enhancement of the quality of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Crane, R.K. Automatic Cell Detection and Tracking. *IEEE Trans. Geosci. Electron.* **1979**, *17*, 250–262. [CrossRef]
2. Johnson, J.T.; MacKeen, P.L.; Witt, A.; Mitchell, E.D.W.; Stumpf, G.J.; Eilts, M.D.; Thomas, K.W. The Storm Cell Identification and Tracking Algorithm: An Enhanced WSR-88D Algorithm. *Weather Forecast.* **1998**, *13*, 263–276. [CrossRef]
3. Anna del Moral, A.; Rigo, T.; Llasat, M.C. A Radar-Based Centroid Tracking Algorithm for Severe Weather Surveillance: Identifying Split/Merge Processes in Convective Systems. *Atmos. Res.* **2018**, *213*, 110–120. [CrossRef]
4. Rinehart, R.E.; Garvey, E.T. Three-Dimensional Storm Motion Detection by Conventional Weather Radar. *Nature* **1978**, *273*, 287–289. [CrossRef]
5. Mecklenburg, S.; Joss, J.; Schmid, W. Improving the Nowcasting of Precipitation in an Alpine Region with an Enhanced Radar Echo Tracking Algorithm. *J. Hydrol.* **2000**, *239*, 46–68. [CrossRef]
6. Liang, Q.; Feng, Y.; Deng, W.; Hu, S.; Huang, Y.; Zeng, Q.; Chen, Z. A Composite Approach of Radar Echo Extrapolation Based on TREC Vectors in Combination with Model-Predicted Winds. *Adv. Atmos. Sci.* **2010**, *27*, 1119–1130. [CrossRef]
7. Horn, B.K.; Schunck, B.G. Determining Optical Flow. *Artifi. Intell.* **1981**, *17*, 185–203. [CrossRef]
8. Lucas, B.D.; Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision. In Proceedings of the IJCAI'81: 7th International Joint Conference on Artificial Intelligence, Vancouver, BC, Canada, 24–28 August 1981; Volume 2, pp. 674–679.
9. Weinzaepfel, P.; Revaud, J.; Harchaoui, Z.; Schmid, C. DeepFlow: Large Displacement Optical Flow with Deep Matching. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1385–1392.
10. Ayzel, G.; Heistermann, M.; Winterrath, T. Optical Flow Models as an Open Benchmark for Radar-Based Precipitation Nowcasting (Rainymotion v0. 1). *Geosci. Model Develop.* **2019**, *12*, 1387–1402. [CrossRef]
11. Srivastava, N.; Mansimov, E.; Salakhudinov, R. Unsupervised Learning of Video Representations Using Lstms. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 843–852.
12. Wang, Y.; Jiang, L.; Yang, M.-H.; Li, L.-J.; Long, M.; Fei-Fei, L. Eidetic 3D LSTM: A Model for Video Prediction and Beyond. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018; pp. 1–14.
13. Weissenborn, D.; Täckström, O.; Uszkoreit, J. Scaling Autoregressive Video Models. *arXiv* **2020**, arXiv:1906.02634.
14. Rakhimov, R.; Volkhonskiy, D.; Artemov, A.; Zorin, D.; Burnaev, E. Latent Video Transformer. *arXiv* **2020**, arXiv:2006.10704.
15. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.; Woo, W. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 802–810.
16. Wang, Y.; Long, M.; Wang, J.; Gao, Z.; Yu, P.S. PredRNN: Recurrent Neural Networks for Predictive Learning Using Spatiotemporal LSTMs. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 879–888.
17. Wang, Y.; Zhang, J.; Zhu, H.; Long, M.; Wang, J.; Yu, P.S. Memory in Memory: A Predictive Neural Network for Learning Higher-Order Non-Stationarity From Spatiotemporal Dynamics. In Proceedings of the IEEE/CVF Conference on Computer, Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9154–9162.
18. Wu, H.; Yao, Z.; Wang, J.; Long, M. MotionRNN: A Flexible Model for Video Prediction with Spacetime-Varying Motions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15435–15444.
19. Geng, H.; Wang, T.; Zhuang, X.; Xi, D.; Hu, Z.; Geng, L. GAN-rcLSTM: A Deep Learning Model for Radar Echo Extrapolation. *Atmosphere* **2022**, *13*, 684. [CrossRef]
20. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features With 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
21. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
22. Ayzel, G.; Scheffer, T.; Heistermann, M. RainNet v1. 0: A Convolutional Neural Network for Radar-Based Precipitation Nowcasting. *Geosci. Model Develop.* **2020**, *13*, 2631–2644. [CrossRef]
23. Trebing, K.; Stańczyk, T.; Mehrkanoon, S. SmaAt-UNet: Precipitation Nowcasting Using a Small Attention-UNet Architecture. *Pattern Recognit. Lett.* **2021**, *145*, 178–186. [CrossRef]
24. Han, L.; Liang, H.; Chen, H.; Zhang, W.; Ge, Y. Convective Precipitation Nowcasting Using U-Net Model. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 4103508. [CrossRef]

25. Song, K.; Yang, G.; Wang, Q.; Xu, C.; Liu, J.; Liu, W.; Shi, C.; Wang, Y.; Zhang, G.; Yu, X. Deep Learning Prediction of Incoming Rainfalls: An Operational Service for the City of Beijing China. In Proceedings of the 2019 International Conference on Data Mining Workshops, Beijing, China, 8–11 November 2019; pp. 180–185.
26. Bai, C.; Sun, F.; Zhang, J.; Song, Y.; Chen, S. Rainformer: Features Extraction Balanced Network for Radar-Based Precipitation Nowcasting. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4023305. [CrossRef]
27. Pathak, J.; Subramanian, S.; Harrington, P.; Raja, S.; Chattopadhyay, A.; Mardani, M.; Kurth, T.; Hall, D.; Li, Z.; Azizzadenesheli, K.; et al. FourCastNet: A Global Data-Driven High-Resolution Weather Model Using Adaptive Fourier Neural Operators. *arXiv* **2022**, arXiv:2202.11214.
28. Geng, H.; Wu, F.; Zhuang, X.; Geng, L.; Xie, B.; Shi, Z. The MS-RadarFormer: A Transformer-Based Multi-Scale Deep Learning Model for Radar Echo Extrapolation. *Remote Sens.* **2024**, *16*, 274. [CrossRef]
29. Xu, L.; Lu, W.; Yu, H.; Yao, F.; Sun, X.; Fu, K. SFTformer: A Spatial-Frequency-Temporal Correlation-Decoupling Transformer for Radar Echo Extrapolation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4102415. [CrossRef]
30. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030.
31. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. *arXiv* **2021**, arXiv:2105.05537.
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
33. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video Swin Transformer. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3202–3211.
34. Shen, X.; Meng, K.; Zhang, L.; Zuo, X. A Method of Radar Echo Extrapolation Based on Dilated Convolution and Attention Convolution. *Sci. Rep.* **2022**, *12*, 10572. [CrossRef] [PubMed]
35. Schaefer, J.T. The Critical Success Index as an Indicator of Warning Skill. *Weather Forecast.* **1990**, *5*, 570–575. [CrossRef]
36. Brotzge, J.A.; Nelson, S.E.; Thompson, R.L.; Smith, B.T. Tornado Probability of Detection and Lead Time as a Function of Convective Mode and Environmental Parameters. *Weather Forecast.* **2013**, *28*, 1261–1276. [CrossRef]
37. Hogan, R.J.; Ferro, C.A.; Jolliffe, I.T.; Stephenson, D.B. Equitability Revisited: Why the "Equitable Threat Score" Is Not Equitable. *Weather Forecast.* **2010**, *25*, 710–726. [CrossRef]
38. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]