1. L034 - What does the HRRR acronym stand for?

Response: High Resolution Rapid Refresh (HRRR). We have explained the acronyms in the revised manuscript.

Changes in manuscript:

‘which is superior to High Resolution Rapid Refresh (HRRR) numerical prediction from National Oceanic and Atmospheric Administration (NOAA) when the prediction time is within 6 hours.’

2. L035 - Explain what are the "U-Net" and "Met-Net" methods.

Response: We have explained "U-Net" and "Met-Net" in the revised manuscript.

Changes in the manuscript:

‘U-Net(Ronneberger et al., 2015) is a well-known network designed for image segmentation, and its core is up-sampling, down-sampling and skip connection. It can efficiently achieve high accuracy with a small number of samples.’

‘Sonderby et al. (2020) proposed a Neural Weather Model (NWM) called MetNet that uses axis self-attention (Ho et al.,2019) to discover the weather pattern from radar and satellite data. MetNet can predict the next 8 hours precipitation with a resolution of 1 kilometer in 2-minute intervals.’

3. L038 - Explain what is a TrajGRU model.

Response: We have explained TrajGRU in the revised manuscript.

Changes in the manuscript: 'Furthermore, they apply the same modification to Gated Recurrent Unit (GRU), and notice that convolution is location-invariant and only focus on fixed location because its hyperparameter (kernel size, padding, dilation) is fixed. But in the QPN problem, a specific location of cloud clusters continuously changes over time. Hence, Shi et al. (2016) proposed Trajectory Gated Recurrent Unit (TrajGRU) that use a subnetwork to output a location-variant connection structure before state transitions. The dynamically changed connections help TrajGRU to capture the trajectory of cloud clusters more accurately. '

4. L040 - Some concise comments about the PredRNN++, MIM, and E3D-LSTM networks are necessary.

Response: We elaborate the description of PredRNN++, MIM, and E3D-LSTM networks.

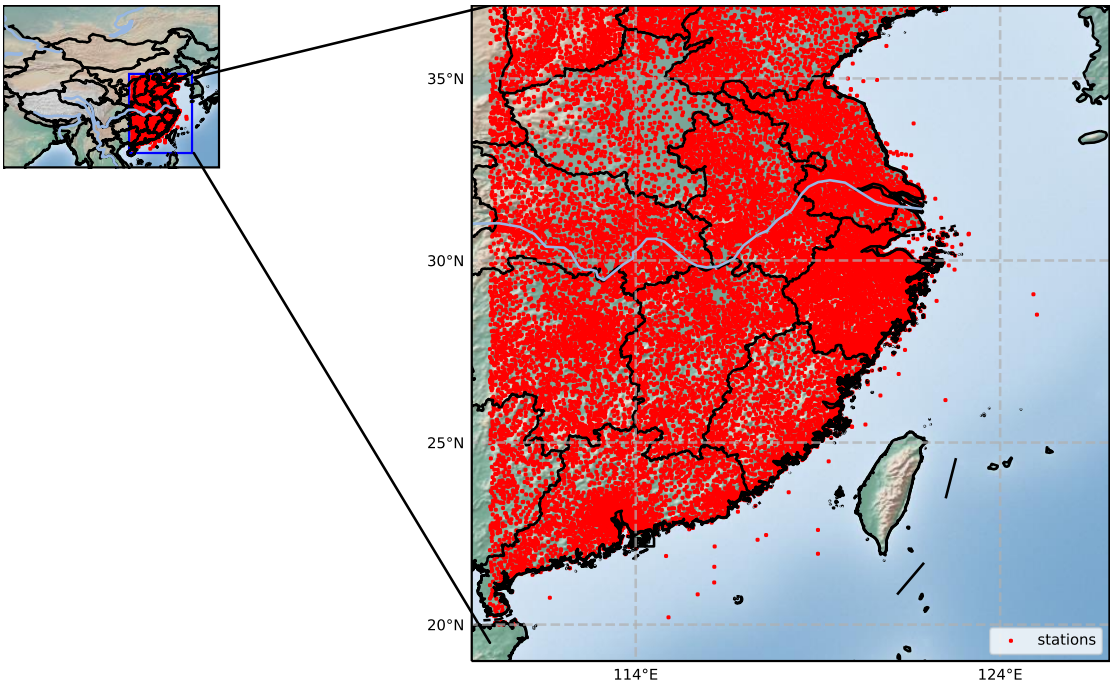
Changes in the manuscript: ‘In the field of video prediction, Wang et al. proposed various recurrent networks based on LSTM. For example, they designed PredRNN++ (Wang et al., 2018) with cascaded dual memory structure and gradient highway unit, which strengthens the power for modelling short-term dynamics and alleviates the vanishing gradient problem respectively. In addition, to capture spatial characteristics through the recurrent state transitions, Wang et.al(2019a) integrate the 3D convolution inside the LSTM units and proposed Eidetic 3D LSTM(E3D-LSTM). Moreover, Wang et al. (2019b) designed Memory in Memory (MIM) to handle higher-order non-stationarity of spatio-temporal data. By using differential signals, MIM can model the non-stationary properties between adjacent recurrent states. However, their work is

based on a slight modification of existing techniques demanding massive computing resource to train and have not been applied to the numerous meteorological data.'

5. L065 - The Figure 1 is not legible. You should improve it.

Response: We have improved it.

Changes in the manuscript: We replace Figure 1 with a legible graph.



6. L122 - I wonder if 240 days of data are enough to train the MSDM. Is this choice explained by a limitation in the computations or is there another justification?

Response: Yes, 240 days may not be enough for training due to the limitation of collecting data. But we make the following justification:

1) We collect 292 days of data and split them into three parts: 80% for training set, 10% for validation set, and 10% for test set. The training set includes several typical types of rainfall events over East China: Convective precipitation, Advection precipitation, Typhoon precipitation.
2) The larger amount of data is to prevent overfitting of the model and enhance the generalization ability of the model. We introduce the early-stopping strategy to monitor the model's performance on validation set to prevent overfitting.
3) U-Net has been proved that it can achieve high accuracy on small number of samples. Therefore, we believe the deep learning part of MSDM based on the modification of U-Net has the same ability.
4) We think that the characteristics of precipitation in a region keep changing over time. The model we trained is based on the data of recent years. Hence it could capture the recent characteristics of the precipitation. Training with long-term data will obtain more general characteristics, while erasing these typical unique characteristics.
5) In the future, we will collect more data to do further research.

7. L 135 - Figure 5 corresponds to a particular date and time. The authors should indicate what they are on the figure. Moreover, I wonder what would be the results for other dates and times. There are too few results presented for the validation and test of the AI methods. More results should be shown.

Response: The date and time of figure 5 is 201809070000. We will add other examples in the revised manuscript.

Changes in manuscript: We add the date and time of figure 5. More examples will be shown in the revised manuscript.

8. L 142 - I do not understand: "it tracks features by the corner detector". What does it mean?

Response: In computer vision, corner (also known as interest points) is the uniquely recognizable characteristics of a image. Corner detector, for example, Harris corner detector, is one of the algorithms for searching these corners. More details will be found at https://docs.opencv.org/3.1.0/d4/d7d/tutorial_harris_detector.html. We will rephrase the sentence to make it easy to comprehend.

Changes in the manuscript: 'However, the fatal weakness of the Optical flow method is that it simply predicts the movement of radar echo from previous images without predicting decay and initiation of radar echo, which causes its accuracy to decrease over time (Table 1) and the false alarm ratio keeps increasing (Table 3). Besides, it employs an algorithm called corner detector (Ayzel et al., 2019) to identify special points from previous frames, and track the movement of these points. When it extrapolates the tail of radar echo, it cannot find corresponding points from previous images (due to the tail of the radar echo at this moment was in a position outside the radar image of previous frames). Consequently, there exist unreasonable shapes in the tail of predicted radar echo.'

9. L 156 - Table 1 - I guess that the Critical Sucess Index is given for four methods, but only for one date and time. What about other test-cases? I think that the methods should be benchmarked in a large number of situations in order to be able to comment the scores.

Response: Table 1 is the average CSI on test set of four models, not a certain day. In the revised manuscript, we choose six thresholds (0.1, 1, 5, 10, 25, 40) and introduce more metrics (HSS,FAR,SSIM) to evaluate model performances.

Changes in the manuscript:

Table 1. Weighted average CSI on test set with different thresholds (0.1, 1, 5, 10, 25, 40, unit: dBZ). The best score is in bold-face. The second-best score is underscored (The greater the better).

Model	**30 min**	**60 min**	**90 min**	**120 min**
Optical Flow	**0.414**	<u>0.303</u>	0.209	0.205
ConvLSTM	0.399	0.269	0.211	0.157
U-Net	0.348	0.259	0.216	0.184
MSDM_mse	0.362	0.286	<u>0.245</u>	**0.218**
MSDM_ssim	<u>0.405</u>	**0.317**	**0.258**	<u>0.217</u>

Table 2. Weighted average HSS on test set with different thresholds (0.1, 1, 5, 10, 25, 40, unit: dBZ). The best score is in bold-face. The second-best score is underscored (The greater the better).

Model	30 min	60 min	90 min	120 min
Optical Flow	0.512	0.409	0.34	**0.304**
ConvLSTM	0.487	0.311	0.246	0.18
U-Net	0423	0.307	0.25	0.209
MSDM_mse	0.437	0.341	0.29	0.255
MSDM_ssim	**0.514**	**0.413**	**0.343**	0.291

Table 3. Average FAR on test set with different thresholds (0.1, 1, 5, 10, 25, 40, unit: dBZ). The best score is in bold-face. The second-best score is underscored (The smaller the better).

Model	30 min	60 min	90 min	120 min
Optical Flow	0.316	0.391	0.439	0.474
ConvLSTM	0.265	0.295	**0.242**	**0.246**
U-Net	0.293	0.309	0.313	0.309
MSDM_mse	0.329	0.364	0.387	0.399
MSDM_ssim	**0.237**	**0.27**	0.303	0.335

10. Otherwise, the MSDM ranks very differently depending on the observation times (from 30 to 120 minutes) with the 0.1 dBZ threshold. Is it logical and explainable? Is it worth noticing that the MSDM ranks consistently (second best score) with the 40 dBZ thresholds. What would be the scores of the MSDM for the Radar Echo Extrapolation at other dates and times?

Response: We choose six thresholds (0.1, 1, 5, 10, 25, 40) and introduce more metrics (HSS,FAR,SSIM) to evaluate model performances. To stress the importance of areas with large radar reflectivity, we assign a weight $w(\text{threshold})$ (Eq. 9) to different thresholds and calculate the weighted CSI and HSS.

Changes in the manuscript:

$$w(\text{threshold}) = \begin{cases} 1, & \text{threshold} = 0.1 \\ 1, & \text{threshold} = 1 \\ 2, & \text{threshold} = 5 \\ 3, & \text{threshold} = 10 \\ 5, & \text{threshold} = 25 \\ 8, & \text{threshold} = 40 \end{cases}, \tag{9}$$

The results of scores have been shown in the previous comment.

11. L 165 -It seems that using the Modified Structural Similarity Index (denoted by SSIM) is counter-productive in terms of MAE and RMSE. Why use it? Once more, I wonder if the example of results produced for one date and time has a general value.

Response: We train each model with MSE loss function to make comparison with the model trained with SSIM. Examples and evaluating scores have been shown in the previous response. Fig 5 shows that MSE cannot predict the large-value area of radar echo.

Changes in the manuscript: We train the MSDM using MSE loss function to make comparison with the model trained with SSIM (Fig 1 and Fig 2). Explanations have been made in the revised manuscript.

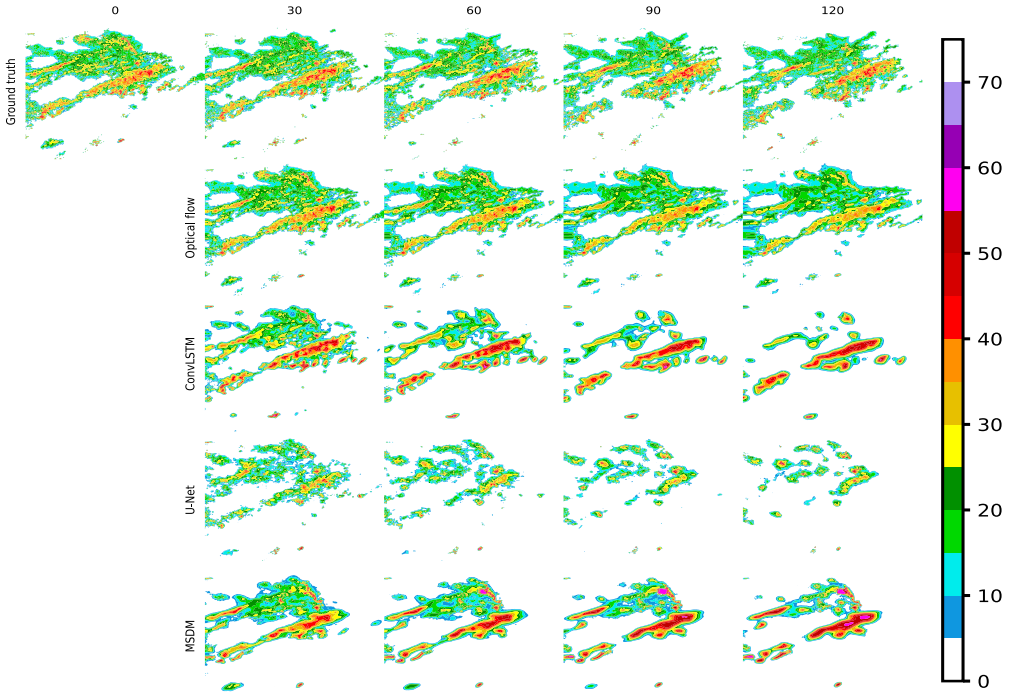


Figure 1 Models trained with SSIM

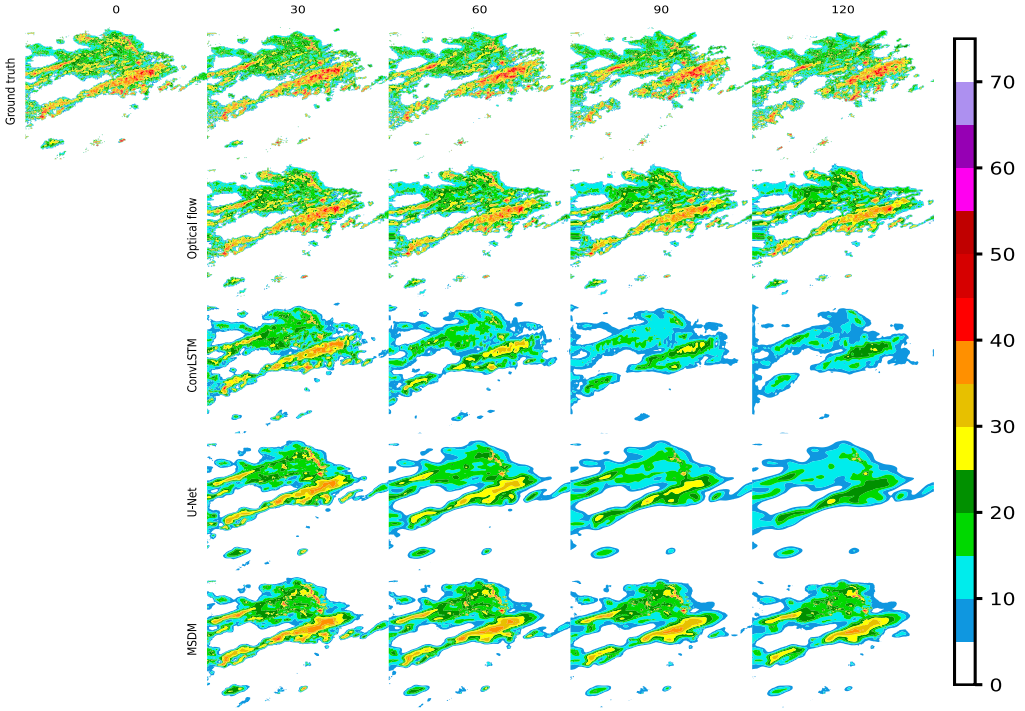
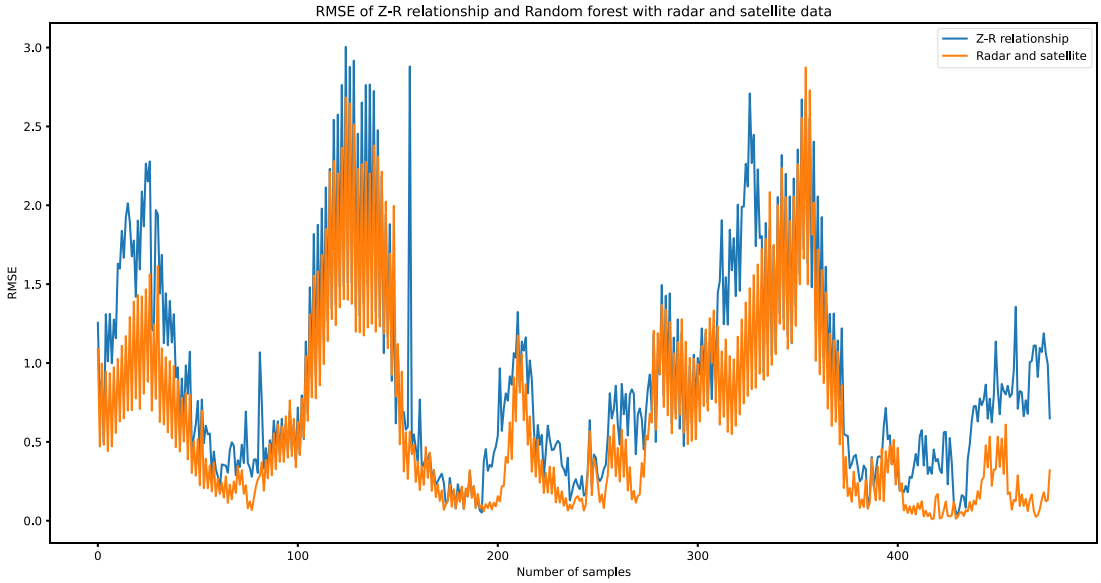


Figure 2 Models trained with MSE

12. L 186 - Figure 8 - Looking at this figure, I am not very convinced that the CSI of the Quantitative Precipitation Nowcasting are better using the random forest than using the Z-R relationship. In general, the scores are quite similar. Could the authors try to better advocate the random forest method?

Response: The CSI describe the spatial distribution of precipitation. We use RMSE and to describe the accuracy of different methods in the revised manuscript. When estimating the precipitation rate at a specific period, the number of data sample is very small, so it is not suitable for methods such as deep learning that require big data. Regressive method such as Random forest will perform better on a small sample of data.

Changes in the manuscript:



13. L 212 - The acronyms "RNN" and "GRU" should be developed.

Response: It has been developed.

Changes in the manuscript: ‘Recurrent Neural Network(RNN), Gated Recurrent Unit(GRU).’

14. My general feeling about the AI methods used separately or combined together through the paper is that all of them have advantages and drawbacks. I suggest to the authors to add a final synthetic table describing the strong points and weak points of the methods and of their combinations. This would greatly help the readers to understand the arguments of the authors.

Response: Thank you very much for this suggestion. We summarize these methods and evaluate them in terms of 12 aspects in the revised manuscript.

Changes in the manuscript: We copy the Table 4 here and the discussion about their advantages and drawbacks has been made in the revised manuscripts.

Table 4. Evaluation on four models with (The less the better)

	The amount of data required for training↓	Time used for training model↓	False Alarm Rate↓	Accumulative system error↓
Optical flow	1	1	2	1
ConvLSTM	4	4	3	2
U-Net	2	2	4	2
MSDM	3	3	1	4

(The more the better)

	The ability to capture spatial characteristics↑	The ability to capture temporal characteristics↑	The ability to predict initiation and decay of radar echo↑	0~1 hour forecast accuracy↑
Optical flow	1	3	1	3
ConvLSTM	2	4	2	1
U-Net	3	1	3	2
MSDM	4	1	4	4

(The more the better)

	1~2 hour forecast accuracy↑	The ability to maintain the shape of radar echo↑	Clarity of radar image↑	Conform to the laws of physics↑
Optical flow	1	4	4	4
ConvLSTM	4	1	1	1
U-Net	3	2	2	2
MSDM	2	3	3	3