1. The critical success indexes (CSI) depends on the artificial threshold chosen by the authors, making it difficult to judge whether MSDM performs better than Optical flow and ConvLSTM. The authors might choose several thresholds (i. e., 0.1, 0.3, 1, 3, 10, 40), and calculate the average CSI of these thresholds, then we can compare these models more easily

Response: Thank you for your advice. 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 manuscript:

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

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	<u>0.512</u>	<u>0.409</u>	<u>0.34</u>	**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**	<u>0.291</u>

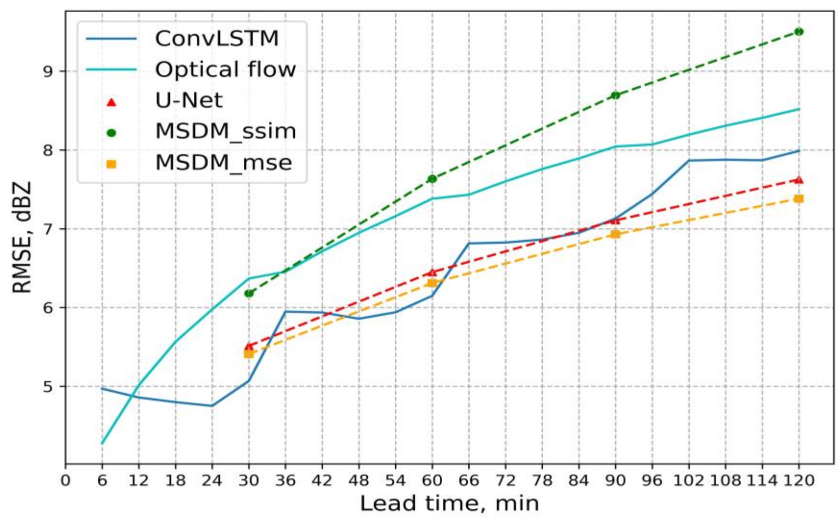
Table 3. Average FAR on test set with different thresholds (0.1, 1, 5, 10, 25, 40, unit: dBZ). The best score is in bold-face. The second-best score is underscored (The smaller the better).

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

2. RMSE is used frequently to judge the performance of machine learning models, but the RMSE of MSDM is too high compared with optical flow, so I suggest the authors to improve and re-train the MSDM model to get a lower RMSE.

Response: We modify the parameters of MSDM and get lower RMSE. Besides, we also trained MSDM with MSE loss function, which has the lowest RMSE. We show the comparison in the revised manuscripts. There are three points to be noted here: 1. MSDM was trained with SSIM loss function, whereas other models were trained with the MSE loss function. Therefore, MSDM is not going to minimize the RMSE but to maximize the structural similarity. 2. RMSE evaluate the global error of predictions. But it cannot evaluate the performance of local area. MSDM gets higher CSI and lower FAR, so it predicts better than other models in local area. 3. We should not focus on one metric. In the revised manuscripts, MSDM outperforms other models in SSIM, CSI, HSS and FAR.

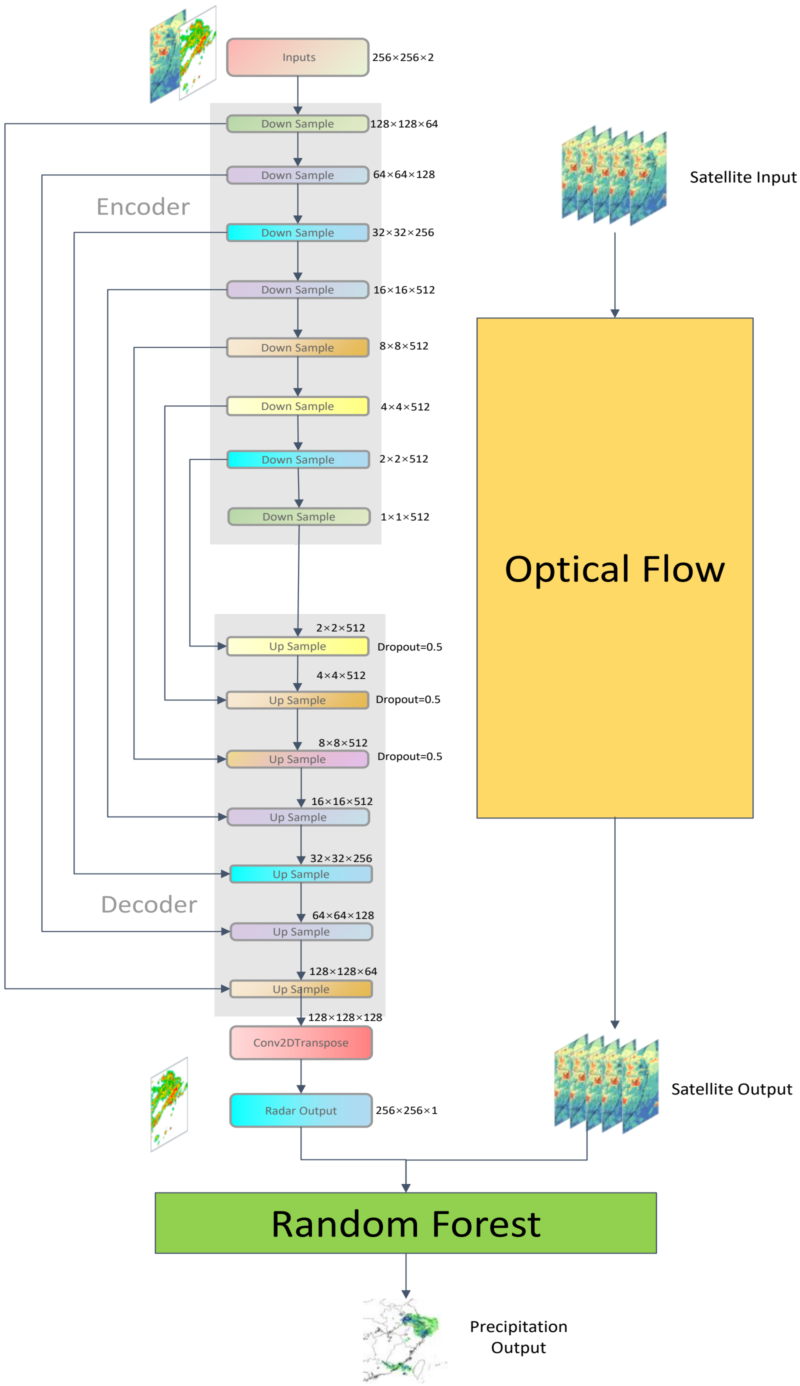
Changes in manuscript: The results of MSDM trained with MSE loss function has been added in the revised manuscripts. It achieves lowest RMSE but performs poor on other metric.



3. The MSDM should be described in more details. In Fig. 3, the red arrow on “ours” (MSDM) indicates that the optical flow is used in MSDM, but in Fig. 4 there is no optical flow in the structure of MSDM?

Response: Fig 3 is a general description of the four models, including the iterative process after one round of predictions. Fig 4 provides more details about the deep learning part of MSDM (feature map, skip connection, convolution etc.). We will put every part of MSDM in Fig 4 in the revised manuscript (including Deep learning, Optical Flow, Random Forest).

Changes in manuscript: We modified Fig 4 and explain it in the new part ‘Model architecture’.



4. Why not predict precipitation directly using MSDM?

Response: We explain the reason why we do not predict precipitation directly using MSDM in the ‘Model architecture’ part.

Changes in manuscript:

‘The reason why we do not predict precipitation directly using deep learning part are as follows: 1) The precipitation data we collected is irregular site data, which is only distributed on land and does not include precipitation on the sea (Figure 1). Whereas the combined radar reflectivity (Figure 2a) and Himawari 8 satellite data (Figure 2b) are regular grid point data and include the data on the sea. The spatial distribution of these three types of data is inconsistent, so it is impossible to make a feature-label correspondence to directly predict precipitation. 2) The use of shapefiles to extract radar echo or satellite data on land will cause the edge of the echo to be limited to the land, which loses the meaning of extrapolation. 3) We hope to improve the transferability of MSDM that can integrate different kinds of data except grid point data. Therefore, the method of processing precipitation data can be used on other observation site data in the daily operation. 4) We believe that deep learning extracts the long-period trend of precipitation efficiently, but it cannot capture the transient characteristics of precipitation. Therefore, for each rainfall event, we use Random forest to model the non-linear relationship between multi-source data to capture its unique characteristics.’