Response to Reviewer #1

REVIEWER # 1

We must thank Reviewer #1 for providing us with useful comments for improving the quality of the paper. We have gone through the comments and have made revisions accordingly.

Comments to the Author
Predictive Network. Through experiments, it can be confirmed that the model has some improvements in middle intensity rainfall prediction. However, there are still some problems need to be solved:

1. The mechanism of RAB and RAM improving the prediction accuracy of middle intensity rainfall rather than other rainfall is not explained.

Reply: Thanks for your question. From tables 3, we can find that the HSS and CSI of those models with RAB and RAM ($RAP-Cell_x$, $RAP-Cell_h$, RAP-Cell) are higher than the PredRNN model in the lowest threshold 5dBZ, which shows that RAB and RAM also can improve the accuracy in other rainfall.

2. The $\hat{X}_2$, $\hat{X}_{T-1}$, $\hat{X}_T$ in Figure 1 is not explained.

Reply: Thanks for pointing it. We have added explanations about $\hat{X}_2$, $\hat{X}_{T-1}$, $\hat{X}_T$ as follows:

"At any timestamp t, model predicts a radar map $\hat{X}_{t+1}$ at the next timestamp t+1 according to the current radar map X_t and historical radar sequence $X_{0:t}$."

3. The authors mention the improvement of RAM parameter performance, but lack discussion of RAB performance. In RAB, the global information is strengthened by self-attention mechanism, which brings more computation and parameters, and improves the prediction accuracy. Compared with the increase of computation, does this increase in accuracy meet expectation?

Reply: Thanks for your question. In the subsection of ablation study, we discuss both of RAB and RAM as follows:

"To investigate the influence of various modules, we conduct an ablation study to discuss the effectiveness of Region Attention Block to the current input and the last hidden state. The result of evaluations is shown in Tables 3 and 4. $RAP-Cell_x$ and $RAP-Cell_h$ denote the PredRNN model embedding the RAB into the input and hidden state, respectively. RAP-Cell model is the combination of $RAP-Cell_x$ and $RAP-Cell_h$, and can also be regarded as RAP-Net without RAM. The results of $RAP-Cell_x$ and $RAP-Cell_h$, are higher than PredRNN, which shows the advantage of introducing Region Attention Block. Specially, the $RAP-Cell_h$, significantly reduces the error according to MAE. Besides, the HSS, CSI and SSIM of RAP-Cell have significant improvements particularly when threshold τ is 40dBZ, which

implies that RAB simultaneously employed in the input and hidden state contributes to the prediction in the heavy rainfall regions. Moreover, by comparing the RAP-Cell and RAP-Net, we find that the RAM can enhance the accuracy of nowcasting especially in the area with middle-intensity rainfall.
Similarly, we also plot Figure 9 to show experimental results of all models against different nowcast lead times. We can see that RAP-Net delivers more promising result when the threshold increases, which demonstrates the effectiveness of combining RAB and RAM in terms of long-term prediction in high reflectivity area. The performance of RAM can be shown by comparing RAP-Cell and RAP-Net. We notice that the introduction of RAM can improve the prediction in the region of middle rainfall intensity. Besides, $RAP-Cell_x$ *and* $RAP-Cell_h$ *embed RAB in the current input and the hidden state, respectively. Their performance is better than the original model PredRNN, especially in 20dBZ threshold. It shows the superiority of RAB. "*

In the application of precipitation nowcasting, the predictions need to be generated within six minutes. Because the observed radar echo map is generated every six minutes. Therefore, the computation only satisfies a requirement that the time cost should be lower than six minutes. As for the proposed model, the generation of predictions only cost 1s. The increase of computation does not cause influence the precipitation nowcasting.

4. In Figure 3, what are the specific advantages of regional attention similarity matrix over other attention?

Reply: Thanks for your question. Figure3 shows three types of attention methods. Traditional attention similarity in Figure3 (a) compares the difference between pixels. The attention similarity from Vision Transformer in Figure3 (b) compares the difference between regions with fixed size and position. The attention similarity from Region Attention (ours) in Figures (c) compares the difference between regions with flexible size and position. Considering that the shape reflectivity of radar echo is irregular and distributed in different places, attention manner in our method can capture the correlation between the different radar echoes better. Therefore, proposed region attention has a better spatiotemporal ability. Besides, we have directly shown specific advantages of regional attention similarity matrix in our manuscript as follows:

"Traditional attention mechanism calculates in Figure 3 (a) the similarity between different pixels and the attention manner in from Vision Transformer in Figure 3 (b) compares different regions in fixed location. Different from both, the attention similarity from Region Attention (ours) in Figure 3 (c) compares the difference between regions with flexible size and position. Due to the irregular shape of radar echo and different distribution, RAB can capture the correlation between the different radar echoes better. Therefore, the introduction of this block can improve the spatiotemporal ability of model.

5. Page 7 Line 128, "$\cdots Q_s \in R^{\mathrm{B\times N\times c*h*w}}$, $K_s \in R^{\mathrm{B\times N\times c*h*w}}$ and $V_s \in R^{\mathrm{B\times N\times c*h*w}}$ are

fed…". What does tensor range mean?

Reply: Thanks for your question. The types of the above tensors (Q_s, K_s, V_s) all belong to real numbers (R), Their size of these tensors all are [$B \times N \times (c * h * w)$] with three dimensions.

6. Please give the specific parameters of RAB and RAM.

Reply: Thanks for your advice, we have offered the specific parameter of RAB and RAM as follows:

"It utilizes four layers RAP-Units as shown in Figure 1 and the parameters setting of each RAP-Unit are shown in Table 1."

Table 1. The parameters setting of RAP-Unit

Attention Type	Name	Kernel	Stride	Pad	Ch I/O	In Res	Out Res	Type
Region Attention Block	CNN_c	5×5	1×1	2×2	64/64	32×32	32×32	Conv
	CNN_{qk}	4×4	4×4	0×0	64/8	8×8	8×8	Conv
	CNN_v	5×5	1×1	2×2	64/64	32×32	32×32	Conv
	Lin_q	-	-	-	-	512	512	Linear
	Lin_k	-	-	-	-	512	512	Linear
Recall Attention Mechanism	CNN	5×5	1×1	2×2	14/64	32×32	32×32	Conv
RNN unit	CNN_x	5×5	1×1	2×2	64/448	32×32	32×32	Conv
	CNN_h	5×5	1×1	2×2	64/256	32×32	32×32	Conv
	CNN_m	5×5	1×1	2×2	64/192	32×32	32×32	Conv
	CNN_o	5×5	1×1	2×2	64/128	32×32	32×32	Conv
	CNN_{last}	1×1	1×1	0×0	128/64	32×32	32×32	Conv

7. Page 8 Line 159. The authors regard X_h^1 as K_c and V_c, then regard H'_tas Q_c. Please explain the reason for this.

Reply: Thanks for your question. The original output of RAP-Cell (H'_t) does not include long-term spatiotemporal representation because its update only utilizes the feature maps at the last time. To extract long-term spatiotemporal representation, we propose Recall Attention Mechanism (RAM) to retrieve the information of X_h^l which convoluted from all inputs sequences X_h^{l-1}. From the equation (7), we can see V_c can be extracted according to the $f(Q_c, K_c^T)$, where Q_c decide how to explore the V_c by dot-producing with K_c. Therefore, by RAM, the original output of RAP-Cell (H'_t) can capture long-term spatiotemporal information. To explain the reason, we have added some explanation in our manuscript as follows:

"From Eq. 7, we can see that the V_c can be extracted according to the $f(Q_c, K_c^T)$, where Q_c decides how to explore the V_c by dot-producing with K_c. Therefore, in l_{th} layer, the

original output H'^{l}_{t} of RAP-Cell can be regarded as query Q_c to explore long-term spatiotemporal representations X^{l}_{h} that is key K_c and value V_c."

8. Page 8 Line 163. "…the new output H_t has recalled all original historical representation and long-term dependencies can be effectively preserved. Besides, the size of the long-memory feature map X_h is fixed at any time." The reason why self-attention can achieve this effect is not explained.

Reply: Thanks for your question. We have added some content to explain why H_t can recall all original historical representation as follows:

"From Eq. 7, we can see that the V_c can be extracted according to the $f(Q_c, K_c^T)$, where Q_c decides how to explore the V_c by dot-producing with K_c. Therefore, in l_{th} layer, the original output H'^{l}_{t} of RAP-Cell can be regarded as query Q_c to explore long-term spatiotemporal representations X^{l}_{h} that is key K_c and value V_c."

Besides, the fixed size of the long-memory feature map X_h at any time has no relationship with self-attention. Because the size of X_h is predefined and corresponding content at different timestamps are fed into X_h. We have added the explanation as follows:

"Because the size of X_h is predefined and corresponding content at different timestamps are fed into X_h"

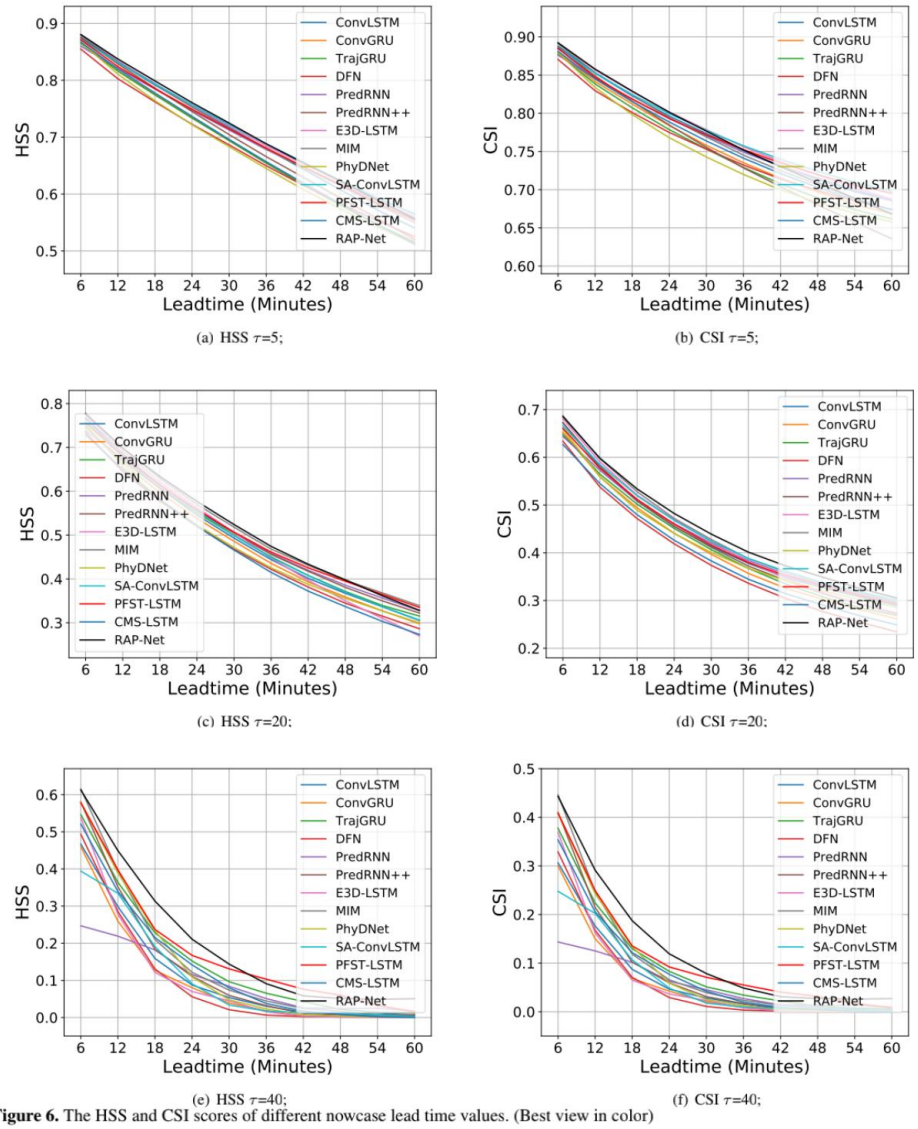
9. As for RAM having more advantages than the recall mechanism of EIDETIC 3D LSTM, the paper lacks important experiment to prove it.

Reply: Thanks for your advice. We have added an experiment about EIDETIC 3D LSTM applied in precipitation nowcasting in Tables 2 as follows::

Table 2. Comparison results on RadarCIKM in terms of HSS, CSI, SSIM, and MAE

Methods	HSS ↑				CSI ↑				MAE ↓	SSIM ↑
	5dBZ	20dBZ	40dBZ	avg	5dBZ	20dBZ	40dBZ	avg		
ConvLSTM Xingjian et al. (2015)	0.7031	0.4857	0.1470	0.4453	0.7663	0.4092	0.0801	0.4186	5.97	0.6334
ConvGRU Shi et al. (2017)	0.6816	0.4827	0.1225	0.4289	0.7522	0.3952	0.0657	0.4043	6.00	0.6338
TrajGRU Shi et al. (2017)	0.6809	0.4945	0.1907	0.4553	0.7466	0.4028	0.1061	0.4185	5.90	0.6424
DFN Jia et al. (2016)	0.6772	0.4719	0.1306	0.4266	0.7489	0.3771	0.0704	0.3988	6.03	0.6268
PredRNN Wang et al. (2017)	0.7082	0.4915	0.1639	0.4606	0.7692	0.4051	0.0901	0.4215	5.42	0.6887
PredRNN++ Wang et al. (2018a)	0.7061	0.5047	0.1710	0.4548	0.7642	0.4176	0.0940	0.4253	5.44	0.6851
E3D-LSTM Wang et al. (2018b)	0.7111	0.4810	0.1361	0.4427	0.7720	0.4060	0.0734	0.4171	5.51	0.6958
MIM Wang et al. (2019)	0.7052	0.5166	0.1858	0.4692	0.7628	0.4279	0.1034	0.4313	5.47	0.6796
PhyDNet Guen and Thome (2020)	0.6741	0.4709	0.1832	0.4427	0.7402	0.4003	0.1017	0.4141	6.25	0.6443
SA-ConvLSTM Lin et al. (2020)	**0.7118**	0.4861	0.1582	0.4520	**0.7725**	0.4161	0.0870	0.4252	5.71	0.6709
PFST-LSTM Luo et al. (2020)	0.7045	0.5071	0.2218	0.4778	0.7680	0.4175	0.1257	0.4371	5.82	0.6367
CMS-LSTM Chai et al. (2021)	0.6835	0.4605	0.1720	0.4387	0.7567	0.3788	0.0948	0.4101	5.95	0.6496
RAP-Net	0.7117	**0.5116**	**0.2293**	**0.4842**	0.7666	**0.4305**	**0.1307**	**0.4426**	**5.37**	**0.7019**

Besides, we have added the result of E3D-LSTM in Figure 6 as follows:



(a) HSS τ=5; (b) CSI τ=5;

(c) HSS τ=20; (d) CSI τ=20;

(e) HSS τ=40; (f) CSI τ=40;

Figure 6. The HSS and CSI scores of different nowcase lead time values. (Best view in color)

Moreover, the new visualization results are shown in Figure 7 as follows:

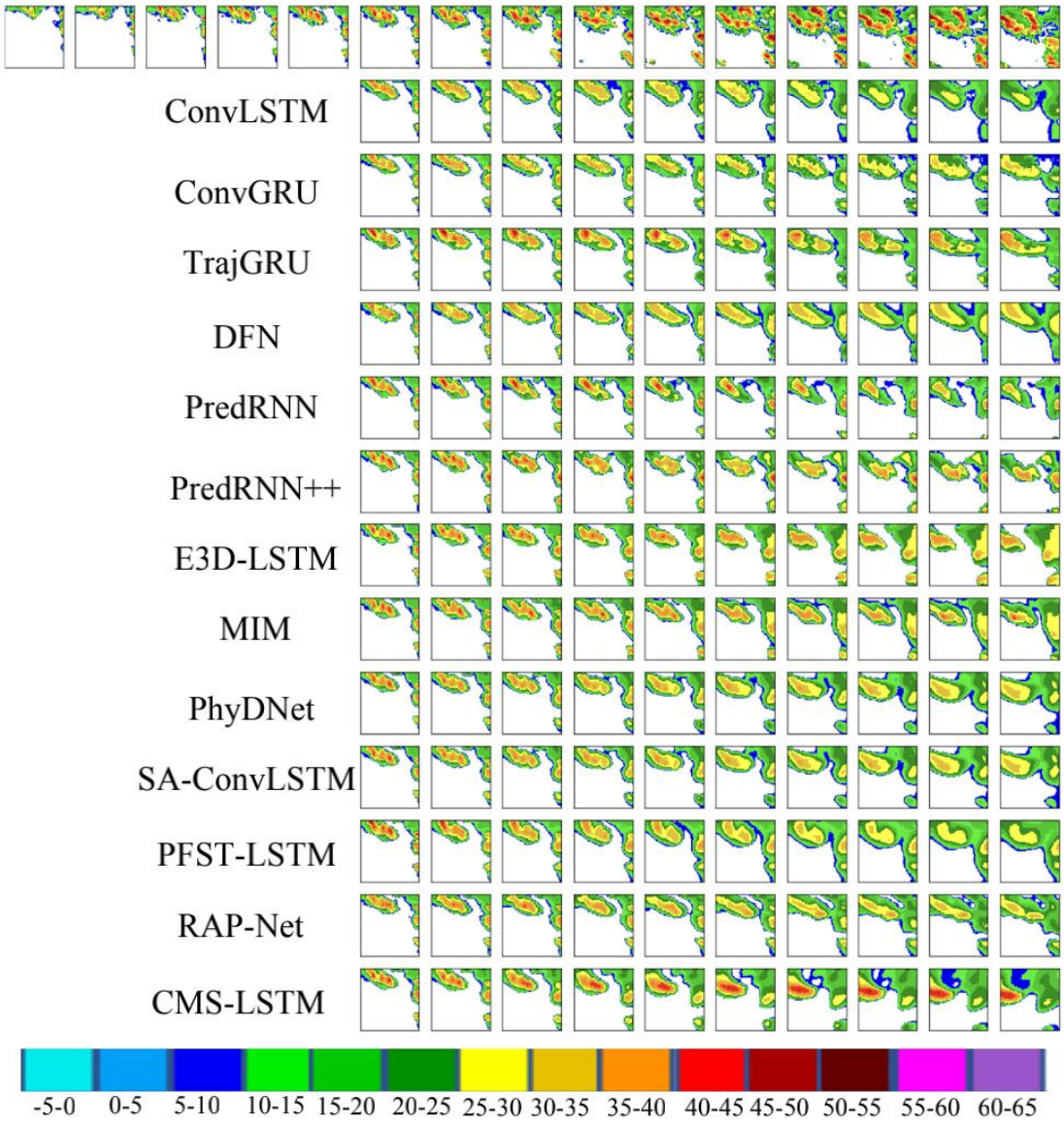


Figure 7. The first row is the ground truth and reminders are the predictions of various methods on an example from the RadarCIKM dataset (Best view in color)

10. What is the resolution of the image processed before experiments?

Reply: Thanks for your question. The resolution of image is 101 x 101. We have added this description as follows:

"Each sequence contains 15 continual observations within 90 minutes, where the spatial and temporal resolution of each map is 101 x 101 and six minutes, respectively."

Minor comments:

1. Page 3 Line 85, "It utilizes the structure of PredRNN …". When the abstract and previous part have been around ConvRNN method, PredRNN is mentioned here. Is it possible to explain the relationship between them?

Reply: Thanks for your question. The ConvRNN is the general name for a series of algorithms that combines convolution and recurrent neural networks. Here PredRNN is a classical method in ConvRNN.

2. Page 11 Line 202, "… which implies the Region Attention can improve …". Is it possible to write RAB and RAM together, not just Region Attention?

Reply: Thanks for your advice. We have modified this sentence as follows:

"Nevertheless its performance is poor in the highest threshold (40dBZ), which implies the RAB and RAM can improve the prediction in the area with high radar echo compared to traditional attention mechanism."