Reply Letter

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 $X_{0:t}$."

The details can be found in line 86 on page 3.

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 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. "

The details can be found in line 233 on page 14..

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.

The details can be found in line 148 on page 8.

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

The details can be found on page 10.

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_h^l that is key K_c and value V_c."

The details can be found in line 168 on page 8.

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'_t of RAP-Cell can be regarded as query Q_c to explore long-term spatiotemporal representations X'_h that is key K_c and value V_c."

The details can be found in line 168 on page 8.

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"

The details can be found in line 172 on page 8.

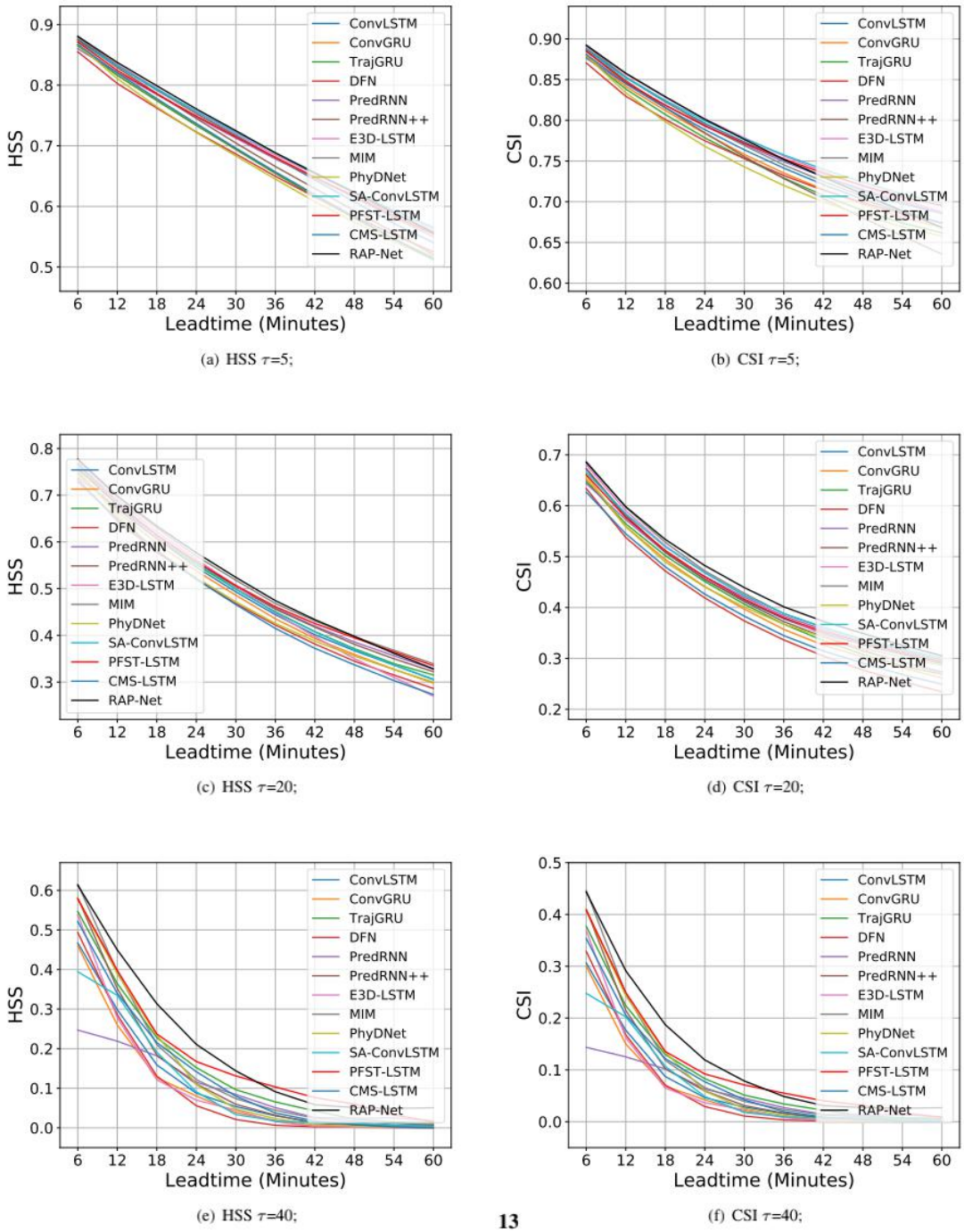
9. As for RAM having more advantages than the recall mechanism of EIDETIC 3D LSTM, the paper lacks important experiment to prove it.

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:

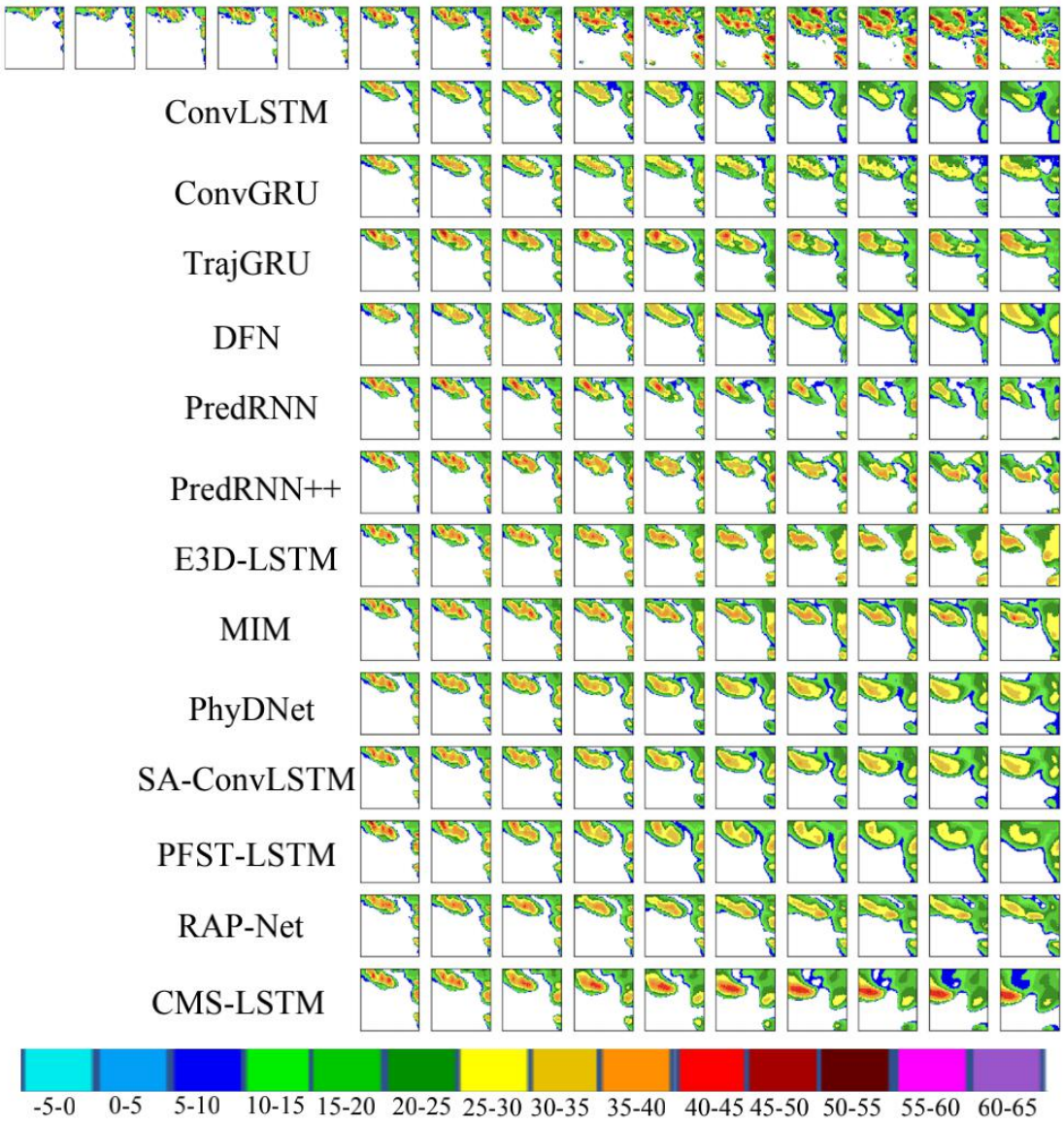


Figure 7. The first row is the ground truth and reminders are the predictions of various methods on an example from the RadarCIKM dataset (Best view in color)

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."

The details can be found in 180 lines on page 9.

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."

The details can be found in line 215 on page 14.

Response to Reviewer #2

REVIEWER # 2

We must thank Reviewer #2 for providing us with useful comments to improve this article. We have gone through the comments and made revisions accordingly.

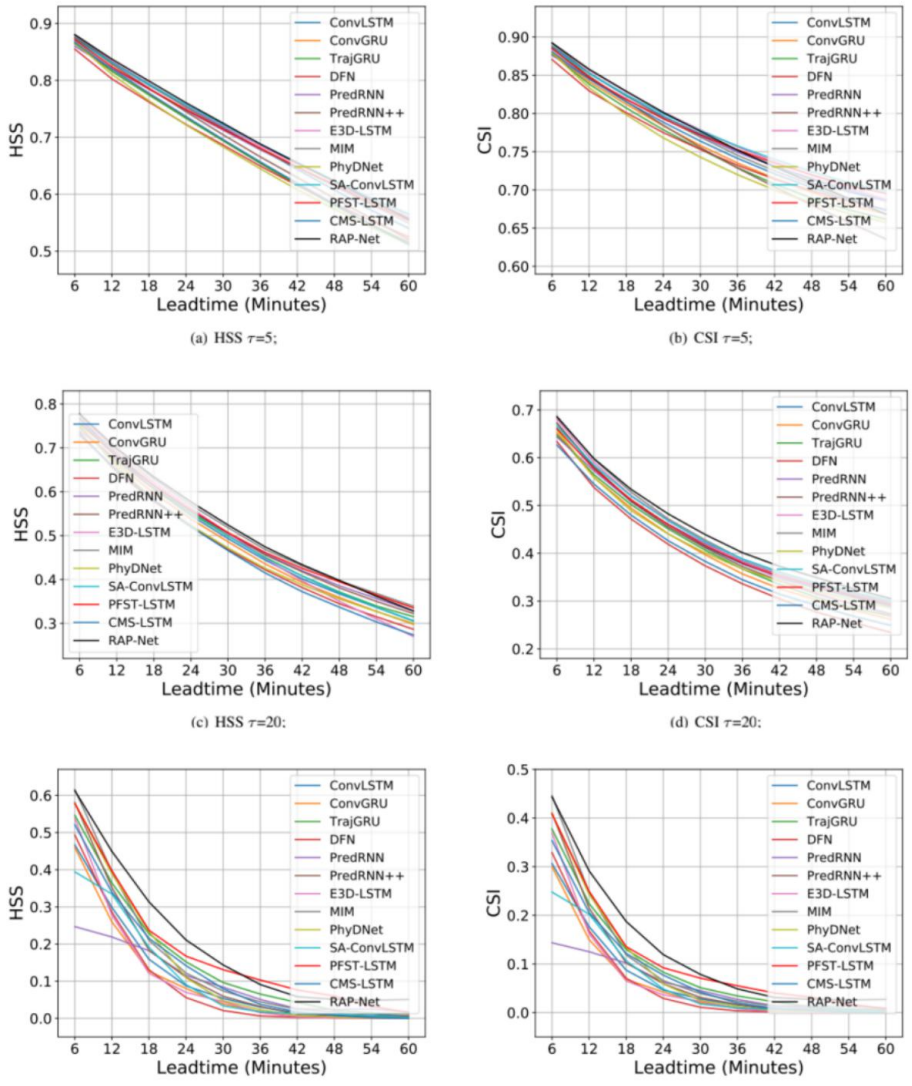
Comments to the Author

The paper describes a RAP—Net network that can be used for radar echo extrapolation. Experiments demonstrate the effectiveness of this method. The authors are suggested to supplement the experimental comparison of high—intensity echoes.

Reply: hanks for your advice. The performance of RAP—Net in high—intensity echoes can be represented from three aspects. Firstly, from the evaluation metrics in Table 2, the HSS and CSI of RAP—Net in the highest thresholds (40dBZ) are higher than other models. Secondly, from Figures 6 (e) and (f), the predictions of RAP—Net keep the best HSS and CSI in 40 dBZ under most of the lead time. Especially, in the last prediction, the HSS and CSI of RAP—Net are obviously higher than other models. Finally, from Figure 7, we can see that only the RAP—Net model predicts color regions. It implies that the proposed model predicts better than other models in heavy rainfall areas. The above three observations jointly confirm that our model is better in high thresholds. The table 2, Figure 6, and Figure 7 can be respectively shown 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**



(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)

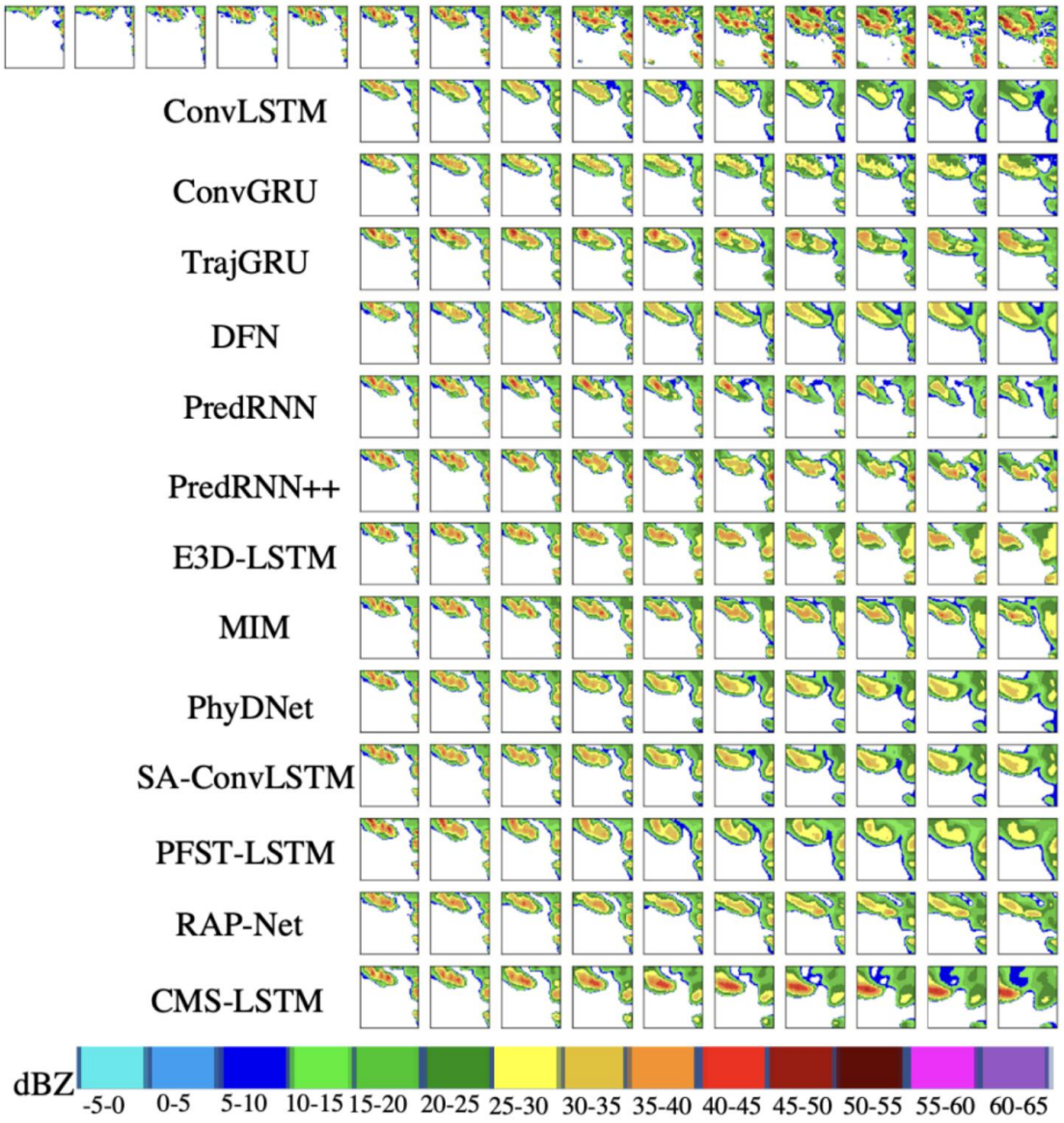


Figure 7. The first row is the reflectivity of ground truth and reminders are the predicted reflectivity of various methods on an example from the RadarCIKM dataset (Best view in color)

Response to Reviewer #3

REVIEWER # 3

We must thank Reviewer #3 for providing us with useful comments to improve this article. We have gone through the comments and made revisions accordingly.

Comments to the Author

To predict weather radar image sequences, the authors propose the model RAP—Net that add attention modules (RAB and RAM) in a ConvRNN model, in order to improve forecasting in the area with heavy rainfall (RAB) and improve the long—term spatiotemporal representation ability (RAM).

1. In general, rainfall intensity is classified as light, moderate and heavy (AMS glossary "Rain": https://glossary.ametsoc.org/wiki/Rain), not as "strong" (line 7) or "middle" (lines 10, 44, 53 and 243). Please fix this throughout the text.

Reply: Thanks for your advice. We have fixed in our manuscript.

2. In line 22, what are the limitations of the traditional approach you are referring to?

Reply: Thanks for your question. As the previous sentence says, the limitation of traditional methods is that they do not exploit abundant historical observations as follows:

"However, these methods do not exploit abundant historical observations."

The details can be found in line 22 on page 1.

3. In line 39, the sentence "where the similar semantics gathered in the same tensor." sounds incomplete.

Reply: Thanks for pointing it. We have rewritten it as follows:

'RAB classifies each feature map into equal—sized tensors and each tensor gatherers a similar semantic."

The details can be found in line 38 on page 2.

4. I suggest reviewing all equations to simplify the notation and unify the letters of all equations (sections 3.2, 3.3 and 3.4):

a. The notation adopted by the terms of equations is confusing (sections 3.2, 3.3 and 3.4):

1. In line 89, what are the letters "l" and "h" in X'_h?

2. Is $X'_h = X_t$?

3. Etc.

Reply: Thanks for your question. Here, the "l" indicates the number of layers. In the bottom layer, X'_h is the result after convolution by historical input sequences $X_{0:t}$. In the other layers, the X'_h is the result after convolution by the X'^{-1}_h. Therefore, the X'_h is not X_t. We have added some explanations in our manuscript as follows:

"Similarly, in the l_{th} layer, the input of the long-memory hidden state is the X'_h. In the bottom layer, X'_h is the result after convolution by historical input sequences $X_{0:t}$. In the other layers, theX'_h is the result after convolution by the X'^{-1}_h. By RAM, the long-term historical representation can be delivered to the next layer."

The details can be found in line 173 on page 8.

b. The equations in lines 109, 100 and 110 do not reflect Fig.2, and vice-versa. The scheme shown in Fig. 2 should also be reviewed with respect to connections.

Reply: Thanks for pointing it. We have corrected errors in equation 2 and Fig.2 as follows:

$$
\begin{aligned}
X'^{l}_t &= RAB(X^l_t),\\
H'^{l}_{t-1} &= RAB(H^l_{t-1}),\\
i_t &= \sigma(W_{xi} * X'^{l}_t + W_{hi} * H'^{l}_{t-1} + b_i),\\
g_t &= tanh(W_{xg} * X'^{l}_t + W_{hg} * H'^{l}_{t-1} + b_g),\\
f_t &= \sigma(W_{xf} * X'^{l}_t + W_{hf} * H'^{l}_{t-1} + b_f),\\
i'_t &= \sigma(W'_{xi} * X'^{l}_t + W_{mi} * M^{l-1}_t + b'_i),\\
g'_t &= tanh(W'_{xg} * X'^{l}_t + W_{mg} * M^{l-1}_t + b'_g),\\
f'_t &= \sigma(W'_{xf} * X'^{l}_t + W_{mf} * M^{l-1}_t + b'_f),\\
C^l_t &= i_t \circ g_t + f_t \circ C^l_{t-1},\\
M^l_t &= i'_t \circ g'_t + f'_t \circ M^{l-1}_t,\\
o_t &= \sigma(W_{xo} * X'^{l}_t + W_{ho} * H'^{l}_{t-1} + W_{co} * C^l_t + W_{mo} * M^l_t + b_o),\\
H^l_t &= o_t \circ tanh(W_{1\times1} * [X'^{l}_t, M^k_t]),\\
H^l_t, X^l_h &= RAM(H^l_t, X^{l-1}_h * W_l),
\end{aligned}
\tag{2}
$$

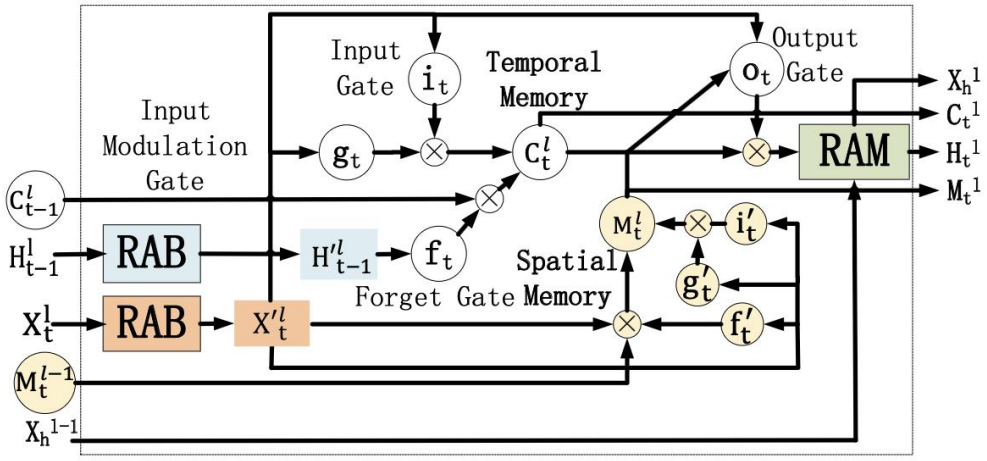


Figure 2. The internal structure of the Region Attention Predictive Unit (RAP-Unit)

The details can be found on page 4.

c. In line 157, when you say "RAP—Cell" do you mean "RAB"? Same in Figure 5, how does "RAM" use "RAP"? (Conflict with Figure 2 and the following equations).

Reply: Thanks for your question. In this paper, we propose two submodules. They are RAB and RAM. By embedding these two submodules, the RAP unit is built as Figure 2 shown. In other words, the RAP unit contains the RAM and RAB. The new equation (2) reflects the calculating process of the RAP unit as Figure 2 shown. The new equation (2) and Figure 2 are shown as follows:

$$
\begin{aligned}
X_t'^l &= RAB(X_t^l),\\
H_{t-1}'^l &= RAB(H_{t-1}^l),\\
i_t &= \sigma(W_{xi} * X_t'^l + W_{hi} * H_{t-1}'^l + b_i),\\
g_t &= tanh(W_{xg} * X_t'^l + W_{hg} * H_{t-1}'^l + b_g),\\
f_t &= \sigma(W_{xf} * X_t'^l + W_{hf} * H_{t-1}'^l + b_f),\\
i_t' &= \sigma(W_{xi}' * X_t'^l + W_{mi} * M_t^{l-1} + b_i'),\\
g_t' &= tanh(W_{xg}' * X_t'^l + W_{mg} * M_t^{l-1} + b_g'),\\
f_t' &= \sigma(W_{xf}' * X_t'^l + W_{mf} * M_t^{l-1} + b_f'),\\
C_t^l &= i_t \circ g_t + f_t \circ C_{t-1}^l,\\
M_t^l &= i_t' \circ g_t' + f_t' \circ M_t^{l-1},\\
o_t &= \sigma(W_{xo} * X_t'^l + W_{ho} * H_{t-1}'^l + W_{co} * C_t^l + W_{mo} * M_t^l + b_o),\\
H_t^l &= o_t \circ tanh(W_{1\times1} * [X_t'^l, M_t^k]),\\
H_t^l, X_h^l &= RAM(H_t^l, X_h^{l-1} * W_l),
\end{aligned}
\tag{2}
$$

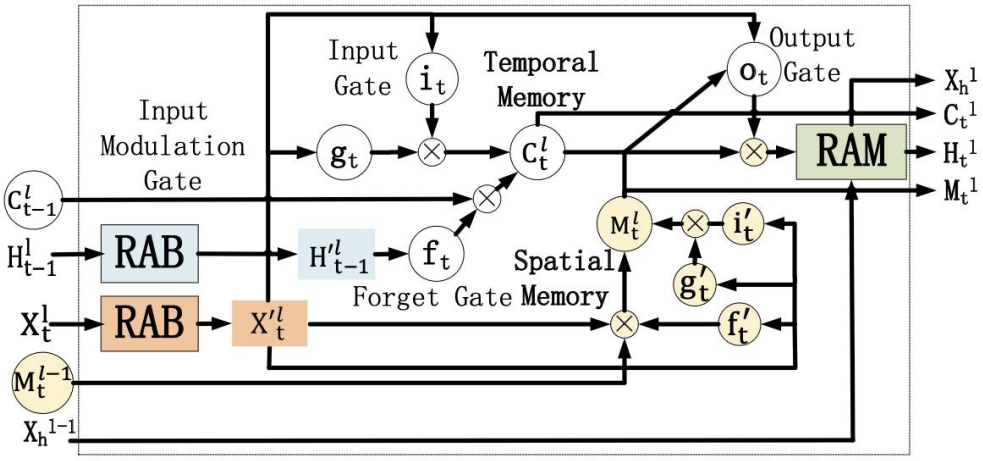


Figure 2. The internal structure of the Region Attention Predictive Unit (RAP-Unit)

Besides, RAP—Cell is the RAP unit without the RAM sub—module. Here, we use Figure 5 to explain how to introduce RAM into the proposed predicted unit. To clear it, we give the explanation of RAP—Cell in the caption of Figure 5 as follows:

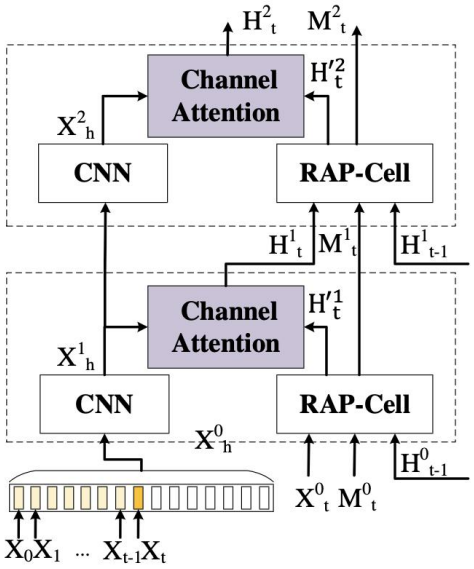


Figure 5. The manner of embedding Recall Attention Mechanism (RAM) into the proposed predicted unit. Here, the RAP-Cell is RAP unit without RAM.

The details can be found on page 9.

5. At the beginning of section 3.3, part of the text is missing, perhaps comments on Figures 3a and 3b.

Reply: Thanks for pointing it. We have added some content about Figures 3 (a) and 3 (b) 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."

The details can be found in line 148 on page 8.

6. In section 4.1, please describe the dataset: data type, variable, instrument; rain or reflectivity; time period, etc.?

Reply: Thanks for your advice. The dataset comes from the CIKM AnalytiCup2017 competition. These radar echo maps are collected by Doppler radar equipment. The type of data is an integer and each value "p" in an image has been linearly transformed. Therefore, the reflectivity (dBZ) in each position can be achieved by follows equation:

$$dBZ = p \times \frac{95}{255} - 10$$

Here, each sample contains 15 continual radar echo maps and covers 90 minuts time period. Finally, these data can be available in:
https://tianchi.aliyun.com/competition/entrance/231596/information

We have added these content as follows:

"The dataset is collected from the CIKM AnalytiCup2017 competition. which covers the whole area of Shenzhen city, China. For convenience, we name this public dataset to RadarCIKM. RadarCIKM has a training set and test set with 10,000 and 4,000 sequences, respectively. There are 2,000 sequences randomly sampled from the training set to build the validation set. Each sequence contains 15 continual observations within 90 minutes, where the spatial and temporal resolution of each map is 101 $\times$ 101 and six minutes, respectively. The range of each pixel is from 0 to 255 and each pixel denotes 1km x 1km. Moreover, the pixel value can be converted to radar reflectivity (dBZ) by the following equation:

$$dBZ = p \times \frac{95}{255} - 10$$

Then, the rainfall intensity can be obtained by the radar reflectivity (dBZ) and Z–R relationship:

$$dBZ = 10 \log a + 10b \log R$$

where the R is rain-rate level, a=58.53, b=1.56."

The details can be found in line 178 on page 9.

7. The equation 8 (page 10) is not a Z–R relation; it looks like a scale conversion. The Z–R relation is in the form of a power law $z = aR^b$.

Reply: Thanks for pointing this mistake, we have corrected this error as follows:

"Moreover, the pixel value can be converted to radar reflectivity (dBZ) by the following equation:

$$dBZ = p \times \frac{95}{255} - 10$$

Then, the rainfall intensity can be obtained by the radar reflectivity (dBZ) and Z–R relationship:

$$dBZ = 10 \log a + 10b \log R$$

where the R is rain-rate level, a=58.53,b=1.56."

The details can be found in line 182 on page 9.

8. In section 4.2, you should inform the range of the evaluation metrics.

Reply: Thanks for your advice. The range of HSS, CSI, and SSIM is [0, 1]. Besides, the range of MAE is [0, ∞]. We have added it in our manuscript as follows:

"Here, the range of HSS, CSI and SSIM is [0,1]. The range of MAE is [0,+∞]."

The details can be found in line 199 on page 11.

9. In line 190, with "SST–LSTM" do you mean "RAP–Net"?

Reply: Thanks for pointing this mistake. The SST–LSTM is RAP–Net, we have corrected it as follows:

"The loss function combines the L1 and L2 to train RAP–Net."

The details can be found in line 204 on page 11.

10. Are the results in Tables 1 to 4 calculated over the entire test set?

Reply: Thanks for your question. Yes, the results in Tables 1 to 4 are calculated over the entire test set.

11. You should consider joining Table 1 with 2 and 3 with 4.

Reply: Thanks for your advice. We have joined the Tables 1 with 2 to Table 2, and Tables 3 with 4 to Table 3 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**

Table 3. Ablation results on RadarCIKM in terms of HSS, CSI, MAE, and SSIM

Methods	HSS ↑				CSI ↑				MAE ↓	SSIM ↑
	5dBZ	20dBZ	40dBZ	avg	5dBZ	20dBZ	40dBZ	avg		
PredRNN	0.7082	0.4915	0.1639	0.4545	0.7692	0.4051	0.0901	0.4215	5.42	0.6887
RAP-Cell$_x$	0.7102	0.5042	0.1754	0.4633	0.7747	0.4235	0.0967	0.4316	5.36	0.6965
RAP-Cell$_h$	0.7149	0.4967	0.1753	0.4623	0.7772	0.4138	0.0967	0.4292	**5.32**	0.7009
RAP-Cell	**0.7234**	0.4757	0.2283	0.4758	**0.7817**	0.4143	0.1300	0.4420	5.64	**0.7036**
RAP-Net	0.7117	**0.5116**	**0.2293**	**0.4842**	0.7666	**0.4305**	**0.1307**	**0.4429**	5.37	0.7019

The details can be found on page 11 and 16.

12. In line 195 "Besides, the proposed model has significant superiority especially for the nowcasting in heavy rainfall regions. Because…" you should join these sentences.

Reply: Thanks for your advice. We have explained some statements by giving the specific experimental results. For example,

"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. Because the main difference between the RAP-Net and SA-ConvLSTM is that they introduce different attention sub modules"

The details can be found in line 215 on page 14.

13. In line 199, what is the execution time of each compared model: PredRNN, PredRNN++ and RAP-Net?

Reply: Thanks for your question. The execution time of each compared model such as PredRNN, PredRNN++, and RAP-Net is several seconds, which is enough satisfy the time requirement of precipitation nowcasting. Because the interval between radar echo images at adjacent times is

6 minutes, which means, as for any model, predictions of all models all can be generated before inputting the next radar echo map.

14. In line 215, please give more arguments to state that your model is better in high thresholds .

Reply: Thanks for your suggestion. The performance of RAP-Net in high thresholds can be represented from three aspects. Firstly, from the evaluation metrics in Table 2, the HSS and CSI of RAP-Net in the highest thresholds (40dBZ) are higher than other models. Secondly, from Figures 6 (e) and (f), the predictions of RAP-Net keep the best HSS and CSI in 40 dBZ under most of the lead time. Especially, in the last prediction, the HSS and CSI of RAP-Net are obviously higher than other models. Finally, from Figure 7, we can see that only the RAP-Net model predicts color regions. It implies that the proposed model predicts better than other models in heavy rainfall areas. The above three observations jointly confirm that our model is better in high thresholds.

15. In Figures 7 and 9, what is being shown, rain or reflectivity? Please fill in the caption.

Reply: Thanks for your advice. It shows the reflectivity. We have filled the statement in the caption as follows:

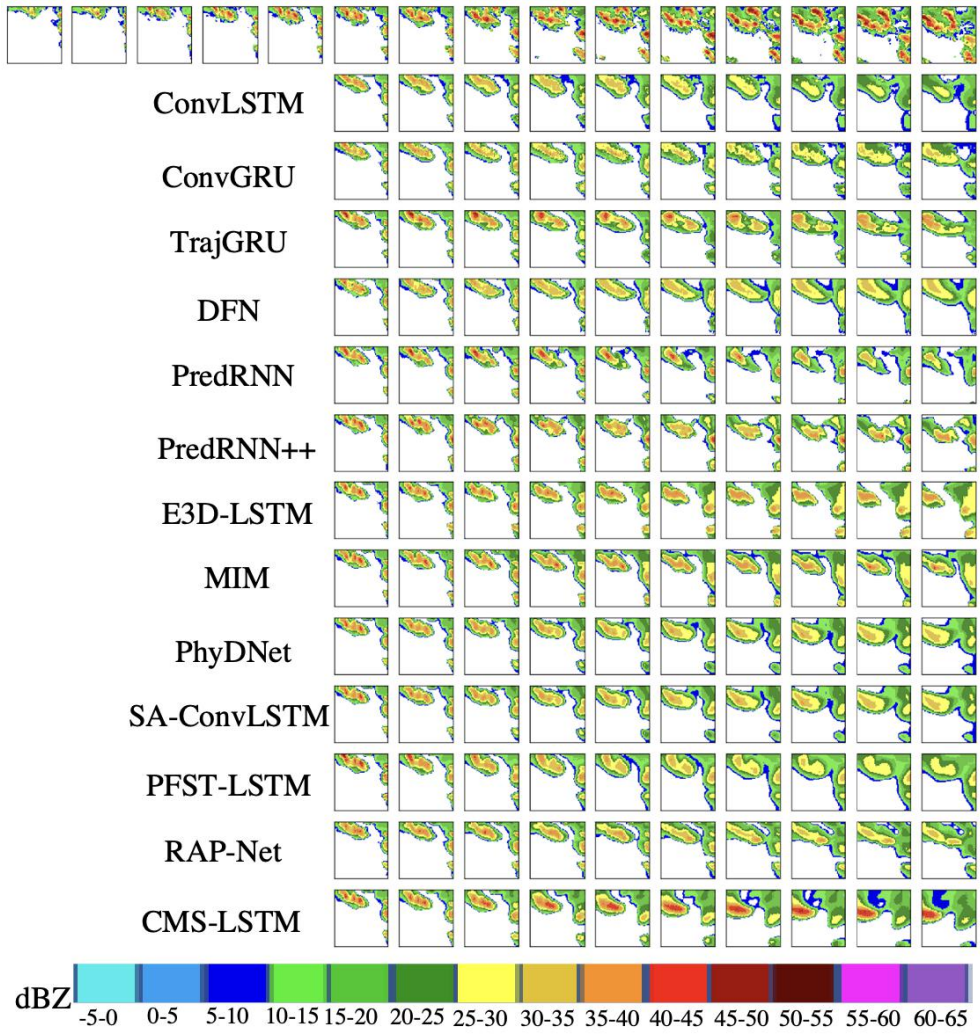


Figure 7. The first row is the reflectivity of ground truth and reminders are the predicted reflectivity of various methods on an example from the RadarCIKM dataset (Best view in color)

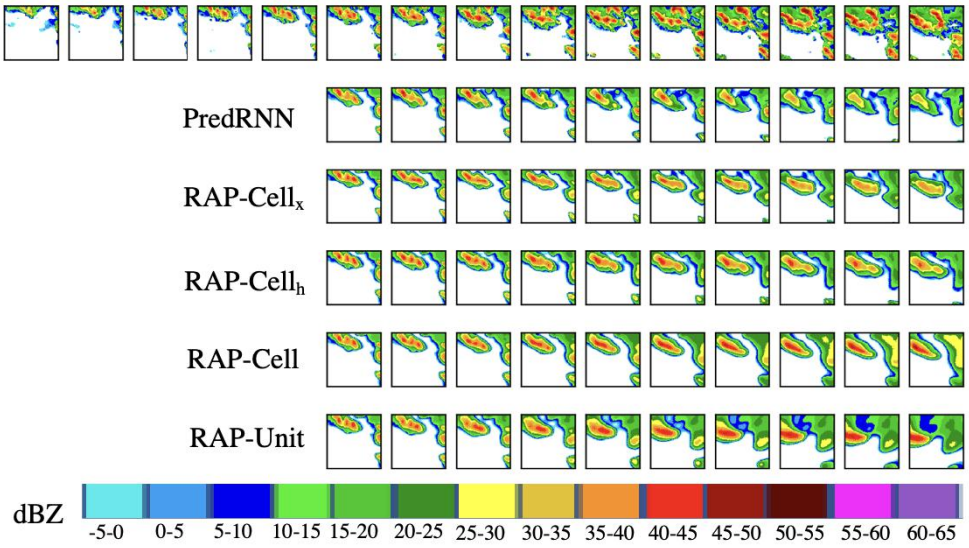


Figure 9. The first row is the reflectivity of ground truth and reminders are the predicted reflectivity of different methods on an example from the RadarCIKM dataset (Best view in color)

The details can be found in page 13 and 16.

16. In the conclusion (section 5), in future work, could you provide more details on how you intend to use more layers? What do you think about computational resources and execution time?

Reply: Thanks for your question. In the conclusion section, we have added some details on how to use more layers to predict radar echo sequence in a single layer as follows:

"Currently, most of existed methods focus on radar echo maps prediction based on a single altitude layer. The variety and movement of echo not only need to consider the previous sequence in the same layers but also need to use different altitude layers. Because the hydrometeors not only happen in the horizontal direction but also act in the vertical direction. For future work, we will consider integrating other layers' historical information to improve the forecasting. In detail, we intend to utilize channel attention to exploit the spatiotemporal representations and then integrate those into the RAP unit. After training, the model can adaptively extract valid spatial information from different levels. We will perform further experiments on multi-channel RAP-Net based on multi-layers radar echo images. Besides, by visualization of the similarity matrix in channel attention, which level is more important for final predictions can be found out."

The details can be found in line 262 on page 17.

Besides, the problem of computational resources is not urgent to be solved. Because the existing hardware is enough to support the computational consumption of the above algorithms. As for large-scale radar echo map predictions, we might introduce some model compression technologies and save more computational resources. Moreover, for the deep learning-based methods, the execution time is often short, which is enough to satisfy the time requirement for precipitation nowcasting. Therefore, we do not add these contents in the conclusion section.