# **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).

 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."

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."

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 "I" and "h" in  $X_h^l$ ?

2. Is  $X_h^l = X_t$ ?

3. Etc.

*Reply: Thanks for your question. Here, the "I" indicates the number of layers. In the bottom layer,*  $X_h^l$  *is the result after convolution by historical input sequences*  $X_{0:t}$ *. In the other layers, the*  $X_h^l$  *is the result after convolution by the*  $X_h^{l-1}$ *. Therefore, the*  $X_h^l$  *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  $X_h^l$ . In the bottom layer,  $X_h^l$  is the result after convolution by historical input sequences  $X_{0:t}$ . In the other layers, the  $X_h^l$  is the result after convolution by the  $X_h^{l-1}$ . By RAM, the long-term historical representation can be delivered to the next layer."

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{split} 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}' + W_{co} * C_t^l + W_{mo} * M_t^l + b_o), \\ H_t^l &= o_t \circ tanh(W_{1 \times 1} * [X_t'^l, M_t^k]), \\ H_t^l, X_h^l &= RAM(H_t^l, X_h^{l-1} * W_l), \end{split}
```





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

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



(2)



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.

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."

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."

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."

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, \infty]. 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,+∞]."

#### 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."

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 RadarCIRIVI in terms of 1155, C51, 5511VI, and WIA |
|---|
|---|

| 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        | <u>5.42</u> | 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        | <u>0.7720</u> | 0.4060        | 0.0734        | 0.4171        | 5.51        | <u>0.6958</u> |
| 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 | <u>0.4778</u> | 0.7680        | <u>0.4175</u> | <u>0.1257</u> | <u>0.4371</u> | 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 ↑  |        |        |        |        | CS     | 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 |

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"

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)

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."

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.