Review of "Key factors for quantitative precipitation nowcasting using ground weather radar data based on deep learning"

This paper compares different ML configurations for precipitation nowcasting. All the ML algorithms compared are deep learning (DL) algorithms. The aim of the paper is to figure out the best configuration and provide some guidelines for the best modeling strategy. I think this goal project is somewhat ill-defined just because there are so many different choices available, and the results of the study are quite inconclusive (discussed below in specific comments); but still it is important to explore these choices systematically as done in the paper.

There are some major issues with the methodology followed in the paper. The DL models trained using the ADAM optimizer can show very different performances based on random initialization – this is what I have seen in the streamflow simulation applications. Therefore, many different models must be trained using different initializations (i.e., different random seeds); an average of the models can be taken as the final model. It does not seem like the authors have followed this procedure; in which case, the results reported are unlikely to be robust. So, this needs to be taken care of before the paper can be considered for publication.

Further, in the multiple prediction case (MU), one expects that the performance will be worse than the single prediction scenario because in the former the model tries to minimize prediction errors at multiple time-steps. But the advantage of using MU design is that it is trying to capture more information from the data and the synergy between the information at different future time-steps helps improve the prediction. I wonder if it is possible to improve the predictions from the MU by adding small layers at the head of the MU where each layer will be dedicated to a single time-step. This can also be seen as the post-processing of MU predictions. This might be helpful in reducing some of the biases that are shown in Table 4.

The authors have compared their results with the 'persistence' prediction. First of all, persistence has not been defined anywhere in the paper. Assuming that it implies that the forecasted value is the same as at the previous time-step, the reported results are not much better than the persistence (both CSI and MAE). Which raises the question if the DL models are actually extracting any significant information from the data? In this regard, why stop at the previous 120 minutes of past data as input? Why not go further in the past; ConvLSTM could extract useful information from a long past sequence.

Specific comments:

Line 70: Why are RNNs unstable? Explain.

Line 88: Do you want to say the 'sequence length' and 'forecasting length'? Amount of data is usually reserved for sample size.

Line 131: A 3x3 kernel was used in this study. Why? Were other kernel sizes tried?

Line 148: Vanilla RNNs also have 'exploding gradient' problem.

Line 149: Do you want to say increasing sequence length?

Line 161: Why is more diverse input sequence a problem? Yes, there is tradeoff between the calibration sample size and the input sequence length. But did you test what is the optimal sequence length for your models given the amount of available training data?

Section 3.2: What learning rates were used to train the models?

Equations 2 and 3: Missing $1/n$.

Table 5: Looking at these results, it seems that that there is no clear winner among these models. It really depends upon what we want to achieve (the performance metric used and the time-step in in future where we need the forecast). This is why I say that this project is a bit ambiguous. In fact, the main conclusion seems to be that there cannot be any specific guidelines for developing these models. I suspect the if the authors compute their goodness measures (CSI, MSE, etc.) separately for different region of the South Korea; they might find that different models perform better in different regions. These arguments apply to high rainfall case of Table 7 also.

Figures 5 and 6: Use a different color scheme to differentiate between the models? It is a bit difficult to differentiate between them.

Lines 294-296: These two sentences are not logically connected. Basically, the first part of the second sentence 'The two models produced identical 10-minute forecasts' can be removed.

Line 303: Language typo: gap was reduced with 5 mm/h threshold?

Line 306-307: This sentence is unclear. Rewrite.

Line 312: But even the CSI were not much better than persistence. So, does this sentence really make sense?

Lines 316-324; I am not sure the ConvLSTM has been designed properly. Basically, the sequence length of 12 may not be good enough for convLSTM.

I think it still might be a good idea to include MAPLE results in Tables 5 and 7 even though these are based on different datasets. This would tell us at least something how a physics based model performs in comparison to the DL models.

Lines 403-405: This is not really supported from Figure 11-14.

Lines 427-429: This is not really true. As I have mentioned earlier, ConvLSTM may perform better with larger sequence length.

Lines 447-448: This has not been mentioned earlier. Explain.