

General statement

We would like to thank the editor for coordinating the review of our work and the peer-reviewers for their valuable comments on our study. In the following, we will address the referees' comments and present our plans and ideas for revising the manuscript. For clarity, our responses are highlighted in red.

Referee comment #2

The paper presents an interesting new method for nowcasting of precipitation, as well as a new precipitation dataset which could be useful for future machine learning applications. The work is thorough and well thought out, and would be suitable for publication in GMD after a few modifications and clarification listed below:

Major Comments

1. There are a couple of instances where the paper would benefit from an additional grammar check. This is especially noticeable in the abstract, the first few paragraphs of the introduction and the conclusion. I have highlighted some specific examples in my minor comments, but I would advise the authors to check the prepositions they use throughout the paper.

Thanks for the comment. We will carefully go through the whole paper and check the grammar issue. The revision will be traceable in the manuscript.

2. I find it odd that in Figure 3, the models cannot outperform the persistence forecast. Surely the ConvLSTM should be able to at least match the persistence as in effect one of its inputs is the persistence forecast. Do you have any intuition as to why you cannot beat persistence?

Thanks for the question. One possible reason why the models can barely beat the persistence forecast in the first 10 minutes is that the precipitation systems are relatively invariant within this very short time period. In our case, the Eulerian persistence forecast is the latest observations available, hence, which is highly correlated to the ground truth at short lead times. With lead times increasing, its performance degrades quickly. On the other hand, the models generate the forecasts by learning the temporal-spatial dependence from the sequential data, which could include fake textures or noise. It makes the models perform slightly worse than the Eulerian persistence at the first few time steps. But for the longer lead times, all the models are remarkably superior to the persistence.

3. In Figure 3c) and 3d) there is a very large degradation in skill at the 20 minute lead time for the ConvLSTM. As far as I can tell this is not discussed in the work and I think it needs to be discussed as it is quite a stark difference.

Thanks a lot for pointing it out. We will add the explanation in the Result section. One reason of the big difference is that the model performance is evaluated with the skill scores, which is affected by the choice of the reference model, rather than the original score values. Here, we choose the Eulerian persistence as our reference model and evaluate the CSI and ETS with a fairly

high threshold (8 mm/h). For the first time step (lead time of 10 minutes), both ConvLSTM and Eulerian persistence can capture strong precipitation events where ConvLSTM is even better. However, ConvLSTM models are prone to produce blurry predictions in an autoregressive prediction task, where the errors in the prior forecasts are inherited to the later ones. It shows that ConvLSTM becomes less efficient in capturing the strong precipitation events at the second time step while the Eulerian reference forecast still performs well due to the short lead time of 20 minutes. With increasing lead time, existing strong precipitation patterns in the Eulerian persistence get more and more displaced with respect to the ground truth data. Thus, the skill of the persistence forecast drops quickly (CSI and ETS are grid point-level metrics) and ConvLSTM again outperforms the persistence forecast, with positive skill scores.

4. Lines 286-288: It would be nice to see some discussion here as to why CLGAN is superior in terms of dichotomous forecast scores but not for RMSE. What attributes does it have or do other models not have, which help here?

Thanks for the comment. We will explain more about the results in this section. Indeed, the reason why the proposed CLGAN is superior in terms of dichotomous forecast scores, but not for RMSE is discussed in the ablation study (Line 339-342), where we found that the GAN-component encourages the model to generate forecasts which share similar distribution with the ground truth data, rather than just reducing the averaged pointwise loss. Hence, more heavy precipitation events are predicted by the CLGAN model which improves the dichotomous forecast scores. However, more predicted high-value precipitation could cause larger biases, compared to the models only generating low-value forecasts. The problem is magnified with the use of the point-by-point scores, i.e. RMSE, which suffer from the double penalty issue as we point out in Line 235-238.

5. Line 329: You say that CLGAN is doing better than ConvLSTM in the difference plot but it is different to see this in the eyeball norm. It would be useful to have some metrics even if it is just RMSE. Also please clarify what you mean by “fine cells”

Thanks for the comment. We will further evaluate the cases with quantitative metrics and add more details to the description in this section. The term "fine cells" here refers to small structures with high spatial resolutions. These small structures in the CLGAN forecasts indicate that CLGAN can generate more details of the precipitation system.

6. Line 331: The topic change here is very confusing as I thought you were still talking about the case study. Maybe add a sub-section title

Thanks for pointing it out. We will revise this part and make it more clear.

7. Lines 370-371 “It shows...”: This sentence is very unclear. Have you already tried adding additional predictors? If so please provide a reference to this work. If this is future work then the c needs to be rephrased because currently it reads like this is a conclusion of the paper

Thanks for the comment and sorry for the confusion. The use of additional predictors is still

subject to the future work. We will rephrase it, i.e. “The idea is appealing since literature shows that the corresponding predictors and physical constraints can significantly improve the simulation of the targeted variable (Daw et al., 2017; Gong et al., 2022) and including additional predictors is subject to our future work.”

8. Lines 371-372 “In addition...”: Please explain further what you mean by expanding to probabilistic forecasting by adding random noise and how you could do this with your model.

Thanks for the comment. We will add more details about how to do the probabilistic forecasting with the proposed CLGAN model in this section. Indeed, GANs have been successfully developed for a probabilistic framework (Ravuri et al., 2021), by adding random noise (as perturbation) to the inputs which enables the generation of ensembles.

Minor Comments

1. Lines 6-8: This sentence is very difficult to understand and needs to be rephrased.

Thanks for pointing it out. We will revise the sentence as follows: “An efficient optical flow model, DenseRotation, as well as a shallow video prediction model, ConvLSTM, and an advanced one, PredRNN-v2, is performed as the competitors. A series of evaluation metrics, i.e. root mean square error, critical success index, fractions skill score and object-based diagnostic evaluation are used for a comprehensive comparison. We show that CLGAN outperforms the competitors in terms of scores for dichotomous events and object-based diagnostics.”

2. Lines 73-74: The last highlight is very difficult to understand and needs to be rephrased

Thanks for the comment. We will revise it: “A sensitivity analysis is performed on the contribution of the GAN-component in generating forecasts with closer statistical properties of the observed precipitation.”

3. Lines 100-102: This sentence needs to be rephrased

Thanks for the comment. we will rephrase it, i.e. “In our paper, the Eulerian persistence model and the DenseRotation model are used to show how well the traditional methods can perform for the precipitation nowcasting task and how much benefit can be further obtained by using DL-based video prediction methods.”

4. Lines 120-121: This sentence needs to be rephrased

Thanks for the comment. We will revise the sentence, i.e. “One of them is the fully convolutional U-Net architecture, which is a U-shaped hierarchical encoder-decoder network with skip connections. The architecture enables abstraction of features on different spatial scales.”

5. Section 3.1.4: It would be helpful to put references to figure 1 in this section because it is very

difficult to follow the CLGAN structure without references to the figure

Thanks for the suggestion. We will add the reference to Figure 1 here and make it easier for readers to follow.

6. Line 242: You are missing the index i in the expression and it is very confusing to have the forecast expression with the observation expression in brackets. You should separate out the two expressions.

Thanks for pointing it out and sorry for the confusion. We will revise the expression and make it more clear.

7. Figure 4: Please clarify in the text and figure caption what the box and whiskers represent

Thanks for the comment. We will add more details in the text and figure caption. Here, the box shows the range of the first quartile (upper) to the third quartile (bottom) of the scores, the whiskers are respectively the 95th percentile (upper) and 5th percentile (bottom) of the scores.

8. Figure 5: Add units to the legend and mention what the legend is in the caption

Thanks for pointing it out. We will add more details in the figure and caption. The metrics presented in Figure 5a-d are all unitless. Figure 5a shows the number of grid points of the observed and predicted precipitation objects, which has no units. Figure 5b-c show the distance between the centroid of the precipitation object and the western boundary (5b) and the southern boundary (5c). The distance is measured by the number of grid points, which still has no units. Figure 5d gives the aspect ratio of the precipitation objects (short side divide long side), which has no units. Figure 5e gives the orientation angle of the precipitation objects, which is the degree of the angle between the precipitation objects and positive x-axis. Positive values mean that the objects have a northeast-southwest orientation whereas negative values mean that the objects are oriented in southeast-northwest direction.

9. Figure 6: It is interesting that ConvLSTM does better at longer lead times here given that it does worse at longer lead times in the metrics. Do you have any intuition on this? Is it just a quirk of the case study you chose?

Thanks for the question. Actually, ConvLSTM is superior to the others in terms of RMSE and correlation coefficient (see in Figure 3a-b), even at the longer lead times. It indicates that ConvLSTM is capable of learning the precipitation patterns, as shown in the case plot (Figure 6). The drawback of ConvLSTM is that it is prone to generate blurry forecasts and not so efficient in capturing strong precipitation events, which is evaluated in terms of CSI and ETS (see in Figure 3c-d). Hence, it is reasonable that the ConvLSTM can perform fairly well in the case study but the proposed CLGAN is able to generate more detailed structures of the precipitation.

10. Line 344: This line needs to be rephrased

Thanks for the comment. We will remove the first sentence and rephrase it, i.e. “A novel architecture CLGAN is proposed in this work ...”

11. Line 358: More potential than what?

Thanks for pointing it out and sorry for the confusion. We will rephrase the sentence, i.e. “Compared to the conventional method, our results indicate that video prediction models with deep neural networks have a better capability to learn abstractions from data which in turn can improve the prediction of complex evolving systems.”

12. Section 5: It would be interesting to have a comment about how you think your model would perform at longer lead times (say 6hrs). Would you still see such good results?

Thanks for the comment. We will add some comments on it in the Discussion section. Indeed, one reason why we focus on the nowcasting up to 2 hours is that the current NWP models have poor performance in this short range forecasting. With the lead time increasing, from 24 hours to 7 days, NWP still serve as the most efficient and reliable method for weather forecasting. Regarding your question, we expect that the model performance degradation will continue for longer lead times so that the skill will become comparable and later inferior to NWP models due to error accumulation in the auto-regressive prediction task. If we increase the lead steps from 10 minutes to 30 minutes, which means we still only have to predict the next 12 frames to obtain the 6-hour forecasts, there's a high chance that the superiority over NWP models can be prolonged. However, this comes at the price of a coarser temporal resolution.

Bibliography

Daw, A., Karpatne, A., Watkins, W. D., Read, J. S., and Kumar, V.: Physics-guided neural networks (pgnn): An application in lake temperature modeling, in: Knowledge-Guided Machine Learning, pp. 353–372, Chapman and Hall/CRC, <https://doi.org/http://dx.doi.org/10.1201/9781003143376-15>, 2017.

Gong, B., Langguth, M., Ji, Y., Mozaffari, A., Stadtler, S., Mache, K., and Schultz, M. G.: Temperature forecasting by deep learning methods, Geoscientific Model Development, <https://doi.org/https://doi.org/10.5194/gmd-2021-430>, 2022.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., thanassiadou, M., Kashem, S., Madge, S., et al.: Skillful Precipitation Nowcasting using Deep Generative Models of Radar, arXiv preprint arXiv:2104.00954, <https://doi.org/https://doi.org/10.1038/s41586-021-03854-z>, 2021.