# Referee report on Choi and Kim (2021), submitted to GMD

## Subject and scope

In their manuscript *"Rad-cGAN v1.0: Radar-based precipitation nowcasting model with conditional Generative Adversarial Networks for multiple domains"*, S. Choi and Y. Kim evaluate the performance of different deep learning designs for precipitation nowcasting. Specifically, they compare a conditional generative adversarial network (cGAN) to two previously published designs, namely ConvLSTM and U-Net. Rad-cGAN combines a U-Net design (as a generative model) with the PatchGAN design from the Pix2Pix model (as a discriminative model). Verification is carried out on a subdomain of the Korean national radar composite. Furthermore, the transferability of the trained network to other subdomains is evaluated based on different transfer learning techniques.

## Overall evaluation

The exploration of deep learning architectures for precipitation nowcasting is gaining more and more attraction. Several studies have been published in recent years which not only suggested various network designs, but also revealed typical weaknesses and challenges of machine learning techniques, as compared to well-established heuristic techniques based on tracking and extrapolation. Further exploration is certainly warranted. Specifically, I very much welcome the investigation of transferability, as included in this study.

But as much as we need such studies, I have several major concerns that should be addressed before this manuscript is considered for publication. This would require some fundamental changes and enhancements of the analysis, hence I recommend major revisions. I will elaborate my concerns in the following.

## Major comments

### Context with Ravuri et al. (2021)

In their study, Choi & Kim suggest a conditional Generative Adversarial Network. A similar approach had been suggested by Ravuri et al. (2021) with good success for the UK. Unfortunately, Choi and Kim do not put their own design in context with the work of Ravuri et al. (2021). It would be helpful to point out, justify and discuss differences in the network design, and resulting implications.

### Spatial verification set-up

The study is about the development and evaluation of different deep learning designs for precipitation nowcasting in Korea. Surprisingly for me, Choi & Kim limit the model verification to arbitrary spatial subsets of their model domain: first and foremost, to the location of the Soyang-gang dam, and second, to the upstream catchment of the dam. While I cannot see

any hydrological justification to predict the precipitation at the dam location itself, I can understand, in the context of dam operation and early warning, the relevance of predicting precipitation for the dam's catchment. However, as the paper is about nowcasting methods, the limitation to the catchment area is unwarranted as it unnecessarily reduces the amount of data that is available for verification. Besides, I do not understand why the authors first compute the verification metrics for each pixel in the catchment separately and then compute the metric's median from this. In my view, using the median improves the resulting metrics specifically for the Rad-cGAN model since it is prone to produce outliers, as the authors state themselves.

### Lack of a competitive benchmark

It has become - and rightly so - the standard in deep learning benchmarking studies to use at least one competitive benchmark model in order to demonstrate the added value of the data-driven models. Several Python libraries have become available in recent years which allow to generate strong benchmark predictions based on tracking (e.g. optical flow) and extrapolation. The authors themselves cite e.g. PySTEPS (Pulkkinen et al., 2019) and rainymotion (Ayzel et al., 2019). I would like to ask the authors to include at least one strong (and open) benchmark from any such library.

### Insufficient metrics

One of the major issues of deep learning models for precipitation nowcasting is that they struggle to predict intense precipitation, and that they introduce spatial smoothing to account for predictive uncertainty. The smoothing effect becomes increasingly pronounced over lead lead time if the model is applied recursively. This important issue needs to be explicitly and extensively addressed in the present study, specifically since Ravuri et al. (2021) appeared to have made substantial progress on that matter. To this end, various de-facto standards have emerged, e.g. to provide skill scores (such as the CSI) for higher intensity thresholds (up to *at least* 10 mm/h), to evaluate the power spectral density (PSD) for various lead times, and to show the fractions skill score over various spatial scales, lead times and intensity thresholds. The use of correlation, RMSE, NSE and CSI for a threshold of only 0.1 mm/h does not meet these standards.

### Transfer learning, hyperparameters

In my opinion, the transfer learning experiment is the most interesting part of this study, yet it requires further attention and analysis. This includes the following aspects:

- In ll. 361 ff., the authors *"[...] infer that by using transfer learning, a model can be successfully developed with different domains, although it does not optimize the hyperparameter to fit the model with the new domain."* The issue of hyperparameter tuning was not addressed in the manuscript before, though. Which hyperparameters were tuned for the Soyang-gang domain, and to which effect? What are the implications for evaluating the transfer learning if you do not analyse the effects of hyperparameter tuning?

- Of course, case 1 provides an important reference for evaluating cases 2 and 3: What does it mean if cases 2 or 3 outperform case 1 in which the model is fully retrained? In addition, I recommend adding another case: the evaluation of the model without any transfer learning, just using the pretrained weights. I think this is important to appreciate the effects of transfer learning.

### Model software and reproducibility

For model description papers, GMD states that *"[...] code must be published on a persistent public archive with a unique identifier for the exact model version described in the paper or uploaded to the supplement"*. Some code is available on a GitHub repository (https://github.com/SuyeonC/Rad-cGAN), yet, this does neither qualify as a persistent public archive nor as a unique identifier. Instead, the published model version requires a persistent DOI. Furthermore, I am not satisfied with the level of reproducibility provided by the GitHub repository: it lacks sufficient documentation (or, strictly speaking, it is not documented *at all*), it lacks the benchmark model implementations (U-Net, ConvLSTM), and it lacks a minimal working example with corresponding data and pre-trained weights. Speaking of data, I could not find any way to download the radar reflectivity composite data samples as pointed out by the authors in the "Code and data availability" section with the provided URL (https://data.kma.go.kr/resources/html/en/ncdci.html). Maybe this works in the Korean version of the website, but this is not sufficient for a study to be published in GMD. Instead, I suggest that the data (or at least samples) are included in another persistent, openly available repository, and that the authors provide sufficient guidance and a working example how to use the data with their code.

### Presentation quality

The presentation quality of the manuscript needs to be improved. This particularly applies to the quality of the figures which all have a very low resolution. For rainfall maps, make sure that you use colormaps that are appropriate for colour-blind people.

## Specific comments

- I think that two statements in the introduction are incorrect: NWP are *not* the standard tool for nowcasting (l. 30), and rainymotion is *not* data-driven (as suggested in l. 35)

- Formatting of equations is odd: it is difficult to separate them from the main text due to the lack of vertical spacing.

- ll. 82 ff: Speculation - topography does not necessarily suggest anything on rainfall patterns (whatever is meant by "rainfall patterns").

- Fig.2: Please add spatial dimensions to the presented data volumes. Furthermore, for the discriminative model (subplot b), it is not clear how PatchGAN's output (34x34) compares to ground truth (pixelwise, averaging, etc.).

- Model description and analysis:

    - The authors stated that the size of the optimised patch is 34x34. However, it is not clear how that patch is clipped from the generated/ground truth image – the output of the discriminative model has a spatial dimension of 32x32 suggesting that there is some overlapping strategy.

    - ll.141-142 state *"...each pixel of the output referred to the probability that the discriminative model determines each patch of the input pair as the real one."* It would be interesting to see the corresponding results on some real examples in the analysis.

    - Based on Goodfellow (NIPS 2016 Tutorial): *"...If both models have sufficient capacity, then the Nash equilibrium of this game corresponds to the G(z) being drawn from the same distribution as the training data, and D(x) = 1/2 for all x."* It would be interesting to see the corresponding results for the discriminative model on some real examples.

- ll. 227 ff.: "Since the precipitation prediction of the model was more accurate, the prediction and observation showed a strong positive linear relationship." - doesn't make sense to me.

- Fig. 6: In order to appreciate the spatial patterns, I would prefer to see the predicted rainfall instead of the bias. If bias plays a role, it should be expressed in adequate verification metrics.

- I don't think that such metrics should be presented with a precision of 4 digits.

- Fig. 3: I do not see the added value of the (right hand) time series panel - there is not much to see and learn when you look at two months of 10 min data.

- I don't see the need for Tab. 3 when you have Fig. 4.

- L. 277: "All models except for persistence performed extremely well" - what is the basis for such a strong statement?

- L. 283-285: "our model performs better than U-Net in predicting peak precipitation [...] prediction accuracy of the ConvLSTM model for maximum precipitation was higher than that of our model" - please confirm these statements by adequate metrics, not from visual inspection

- Ll. 287 ff: what is the basis for normative statements such as "good" and "sufficient"?

- Why use an entirely different presentation format in Fig. 5 to evaluate the performance in the catchments? Anyway, a revised version of the paper should not evaluate model performance for the dam locations and the catchments.