Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC3
https://doi.org/10.5194/gmd-2021-405-RC3, 2022

Comment on gmd-2021-405

Anonymous Referee #3

Referee comment on "Rad-cGAN v1.0: Radar-based precipitation nowcasting model with conditional generative adversarial networks for multiple dam domains" by Suyeon Choi and Yeonjoo Kim, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-405-RC3, 2022

Rad-cGAN v1.0: Radar-based precipitation nowcasting model with conditional Generative Adversarial Networks for multiple domains

Summary:

This paper is concerned with the prediction of precipitation at high temporal and spatial resolution and short lead-time commonly referred to as precipitation nowcasting. The authors explore the use of conditional generative adversarial networks (cGAN) to generate auto-regressive predictions of rain from radar images (1 km resolution) up to 90 min. The authors compare their results to several baselines from the deep learning field on data collected in sub-regions of Korea. They also explore several fine-tuning strategies to transfer the learning from one region to another.

Overall, the paper tackles an interesting problem where traditional methods such as NWP do not perform well. Given the growing availability of observation data, I expect that deep learning approaches will become more popular in this field. Therefore, this study, and especially its exploration of transfer learning, is timely. However, I have several major comments that require clarifications from the authors before the manuscript can be recommended for publication.

Major comments:

- This paper should provide more context on the existing literature and on how it contributes to the field of nowcasting. In particular Ravuri et. al. 2021 also investigates the use of conditional GAN for precipitation nowcasting with extensive (probabilistic) evaluation. Could the authors further discuss how their method differs from Ravuri et. al.?
- One of the main claims of the paper is to outperform other baselines on a new dataset. To support that claim, the authors need to provide stronger evidences in the form of extensive evaluation:
  - Higher thresholds: The case made for nowcasting is to help with risk management. However, CSI is computed based on the 0.1mm/h threshold, which is pretty low to have any practical impact on risk management. Could the authors add evaluation at higher thresholds. For instance, I suggest adding the thresholds used in Ravuri et. al. 2021. While predicting intense events is more challenging, it is a more informative measure of progress for the field of nowcasting.
  - Additional metrics: Previous studies have also established a set of relevant metrics that should be included in this study. For instance, PSD plots at various lead-times as a proxy for blurriness of the prediction, and skill scores over aggregated regions (e.g. 2x2, 4x4, 8x8 pixels). Note that these metrics are likely to pick up on positive characteristics of GAN predictions (e.g. sharp predictions and spatial consistency).
  - Probabilistic evaluation: One of the main advantages of training a generative model, such as a GAN, is the ability to generate multiple predictions and conduct probabilistic evaluation on forecast ensembles. This leads to the following questions:
    - Does Rad-cGAN generate multiple samples?
    - If so, how are deterministic metrics such as CSI computed? Are the multiple samples aggregated into an average before computing the metrics? Is a single sample used for evaluation?
    - If not, why? Given that the motivation for this work is risk management which relies on assessing risks integrated over the full distribution of possible events, this would seem to be of critical importance.
  - Baselines: Please include PySTEPS (publicly available) as a baseline for your evaluation. Note that to make PySTEPS a competitive baseline, it might need to be fed more context than 128x128, as this is an advection method which is going to be more penalized at the boundaries of the prediction than DL methods.
  - Data leakage and meta-optimization: The split between the different datasets needs to be clarified. It is now common practice to divide the available data between training set (for optimization), validation set (for hyper-parameter tuning and best checkpoint selection) and test set (for final evaluation). While the training set and test set are clearly defined, I did not find any reference to the validation set. This begs the questions:
    - On which dataset were the various hyper-parameters tuning done?
    - Was the early stopping metric defined on the training set or the test set?
    - How have the baselines been tuned?
- The second claim of the paper is to successfully employ transfer learning techniques to generalize to different regions. In the machine learning context, transfer learning is often used to transfer knowledge acquired on a large dataset to a similar but smaller

dataset. While reducing the computational cost of training the model is definitely of interest, it is the limited availability of data for the new task (here the new region) that usually justifies doing transfer learning rather than training from scratch. In the context of this paper, it is not clear why transfer learning is useful: there is the same amount of data for the new regions, and most of the data is disregarded at train time (not included in the crop). Several alternative strategies should be included as baselines:

  - Pre-trained model on first region, without fine-tuning (effectively case 2, since optimizing the discriminator without fine-tuning the generator is like using the pretrained prediction model without modification)
  - Pre-trained model on first region, with fine-tuning (generator+discriminator) on the new region (close to case 3)
  - Pre-trained model on random crops from all the data available (the 11 radars), without fine-tuning
  - Pre-trained model on random crops from all the data available (the 11 radars), with fine-tuning on the new region
  - Training model on new region from scratch (case 1)

Additionally, as a methodological demonstration, the authors should also consider running experiment 3b) with varying amount of data from the new region (for instance, using the equivalent of 1, 2 or 3 summers). This would be an informative experiment (and more realistic use case) on the amount of data required to do transfer learning.

Minor comments:

- Why not work with the precipitation itself, rather than reflectivity? The non-linearity may translate in different error amplifications for different precipitation amounts.
- Please provide more information about the datasets, this includes (per data split):
  - Distribution of precipitation amount
  - Number of examples

This is particularly important to verify if there is enough data to make any assessment about certain events (e.g. high intensity precipitation)

- Why just consider the summer? Is there more precipitation during those months?
- How was the test set normalized? Using min max of training set? Or the test set? This is important to make sure there is no data leakage, and the predictions are not using information from the future.
- The transfer learning part is missing a case, fine-tuning both generator and discriminator.
- Please provide all the plots and figures at higher resolution.
- For transfer learning case 2, what is the use of training a discriminator if the generator

is pre-trained and frozen? The predictions will be unaffected by the fine-tuning. It seems to me that Case 2 is equivalent to just applying the model trained on the first region to a new one.

- For evaluation, please include tables at 90 min lead-time rather than 10 min. Could you also include mean quantities. If you keep reporting the median (which is more computationally intensive as you need to keep track of the whole distribution), please include other percentiles (95%, 99%) to give a sense of the uncertainty.
- Line 69. Please clarify what is proposed on top of the cited paper. If using the technique as is, please state "we apply the transfer learning [...]".
- Line 78. Typo.
- Paragraph 2.2.1. This paragraph is hard to read. Could you please rewrite it in a more precise way.
- I would suggest using "discriminator" and "generator" for the two submodels of the GAN rather than "discriminative model", and "generative model" which have a broader meaning.
- Line 95. Do you mean "It consists of a generator (G) that produces the distribution [...]"?
- Line 122. Do you mean "To prevent overfitting"?
- Paragraph 2.2.3. This paragraph is hard to read, in part because the word input is used to refer to two different quantities. Could you please rewrite it in a more precise way.
- Line 136. What does the following mean? "The size of the patch (N) was determined by the structure of the entire discriminative model, and it increased as the model became deeper. We constructed a discriminator model through optimization with a 34 × 34 patch size." How was N decided? Using hyper-parameter optimization? By "constructing a discriminator through optimization", do you mean training the weights of the model or optimizing the architecture?
- Line 257. This statement is a bit misleading. Fine-tuning is not defined by changing the learning rate to 1/10 of the original setup. It typically uses a smaller learning rate to make small adjustments to the weights. The computational savings come from the lower number of training steps for the fine-tuning rather than the lower learning rate.
- Line 314. I am not sure how this shows that baselines are properly tuned and trained.
- Line 342. How come case 2 (generator trained on region 1 and not fine-tuned on new region) performs better than case 1 which trains (from scratch) on the new region with the same amount of data as the pretraining on region 1? This is a very surprising result. Or am I misunderstanding what case 2 is doing?
- Please include a reference to a quantitative metric (plot or table) for any statement about performance.
- Fig 6. Please also include samples of predictions (not just the error) for different lead-time for all the models/baselines.