EGUsphere, author comment AC3
https://doi.org/10.5194/egusphere-2022-440-AC3, 2022

Reply on RC3

Tarek Beutler et al.

Author comment on "Deep Learning Approach Towards Precipitation Nowcasting: Evaluating Regional Extrapolation Capabilities" by Tarek Beutler et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-440-AC3, 2022

We would like to thank the anonymous referee #3 for their extensive constructive comments and suggestions. We are in the process of revising the manuscript with the referee's suggested changes. Point by point answers to the referee's comments can be found further below.

Content comments:

- **2.1 (data). This section lacks some important information:**
  - **Did you do any pre-processing on the images or is 480x480 pixels the original size of the image?**
    The RADOLAN product is a composite of 17 radar stations which cover a bigger area. We cut out a 480x480 km region over central Germany. This information has been added in the revised manuscript.
  - **Did you use reflectivity or precipitation? In case of precipitation, what was the Z-R relation used?**RADOLAN is a reflectivity product. This information is given on page 3, line 67.
- **Is it a CAPPI or PPI? Please, inform the height or sweep elevation;** We will include the information in the text.
- **The used period from 2017 to 2021 comprises only three years: from Apr/2017 to Mar/2021. Make it clear to the reader;** The amount of days is already given on page 3, line 72. To make it more clear, the amount of years was also added in the revised manuscript.
- **Did you use the complete sequence or only selected rain events, as Shi et al. (2017)?**
  Only days with precipitation were selected from the RADOLAN dataset. This information has been added in the revised manuscript.
- **Explain the selection of training, validation and testing sets: summer to test, winter to validate;**
  Roughly the first 10% of frames were chosen for testing, the last 5% of frames for validation.
- **Instead of "German RADOLAN", use "RADOLAN".**
  This has been changed in the revised manuscript.
- **Please, if possible, give more information: What is the weather radar type: band, polarization, Doppler? Where can the reader find this type of information?**
  All DWD weather radars are Doppler radars. This information has been added in the

revised manuscript.

- **Line 120: "*Other finetuning configurations were tested*" The authors should comment more on this;**
  A more detailed breakdown of the other finetuning configurations was added in the revised manuscript.
- **Table 1:**
  - **The first threshold includes rain rate = 0. What do you consider as no rain?**
    The first threshold includes all rain events with a rain rate R larger or equal to 0 and smaller than 0.5.
  - **The RADOLAN column sum more than 100%;**
    This is a mistake due to rounding and has been corrected in the revised manuscript.
  - **How was this table calculated, with the complete sequence of the dataset or with selected rainy cases? Is it the distribution of pixel values in the image set?**
    The table was calculated with the complete sequence of the dataset, which only contains rainy days. The percentages indeed refer to the pixel distribution. This information has been added in the revised manuscript.
  - **Shi et al. (2017) used selected rainfall events. Is the HKO-7 column considering only these events? (You do not need to repeat Shi et al.'s paper, but you should provide enough information for your reader to understand what you are talking about.)**
    The HKO-7 column only considers the selected rainy days. This information has been added in the revised manuscript.
- **Line 138: The authors introduced binary values (0, 1) based on thresholds. They must inform the meaning of the values above and below the thresholds;**
  We believe the anonymous referee is asking to clarify how the binary values are assigned. If a pixel is above or equal to the currently selected threshold *r*, it gets converted to a 1, otherwise to a 0. This clarification has been added in the revised manuscript.
- **Line 145: What "measurements" do you refer?**
  "Measurements" refers to the two errors in Table 2.
- **Line 145: Briefly describe Welch's t-test in sec. 3;**
  A brief description of the t-test has been added in the revised manuscript.
- **4.1: The model predicts 20 images, from 5 to 100 min lead time. For which forecast times are the shown results?**
  The shown results use the average score of all 20 prediction frames. This information has been added in the revised manuscript.
- **2 is equal as Fig. 3, the same pattern, but with different values. The authors should verify that it is correct. If correct, what is the gain of using such metrics, what does this prove? Why not use another metric to explore more information?**
  Both CSI and HSS measure how accurate a prediction is to the ground truth, so both measurements correlate with each other. There is value in using both scores, as the CSI measures how many rain events were predicted correctly, whereas the HSS measures if our predictions are better than if we had made a random prediction. We agree though on using additional metrics to explore more information and have added the False Alarm Rate (FAR) and Probability of Detection (POD) skill scores to the revised manuscript.
- **Section 4.1 (lines 157-163):**
  - **The values are too small to draw conclusions. The authors forced a conclusion mainly with the expression "big increase" (lines 163, 225)**
    We think that an approximate 2% difference for such a low performing threshold is indeed a significant increase. However we agree that the expression "big increase" can be misleading and removed the word "big" from lines 163 and 225 in the revised manuscript.
- **What about the analysis of the result evolution with forecast time?**

Additional statistical analysis regarding prediction time, similar to the qualitative analysis in section 4.2, has been added to the revised manuscript.

- **I suggest including other metrics, such as FAR and POD, and some metric to assess the image quality, since you are using a computer vision method;**
  See our answer to point 8.
- **Lines 165-167: "*We compare the model output for the real truth in the frame of two case studies. Using the RADOLAN data set, we consider frontal systems at 4 May 2017, 14:40:00 UTC and at 12 May 2017, 07:40:00 UTC. These are two exemplary dates of clusters of mainly moderate precipitation crossing Germany, where the data is not part of the training data set of the model.*" The authors should comment on this first of all in sec. 2.3, Experimental setup;**
  Comments on the case study setup have been added to Section 2.3 in the revised manuscript.
- **4.2:**
  - **You compare your results with Ravuri et al. (2021) and Ayzel et al. (2020). How many examples did these studies use to compute their statistics? Because yours considers just one case;** Line 195 we added the information: "Ravuri et al. (2021) consider a single case study (...)" and in Line 210 ff. we changed the text to *Ayzel et al. (2020) (...). The authors select 11 events during the summer months of the verification period (2016–2017) and evaluate the models RainNet and Rainymotion for the intensity thresholds 0.125 mm/h, 5 mm/h and 15 mm/h for prediction time up to 60 min.*
- **Line 199: "the positive effect of finetuning is clearly visible for higher precipitation intensities." You are referring to a small gain of just one score;**
  This line has been changed in the revised manuscript.
- **What is the merit of the model used in your predictions? The other models have different architectures; you should take this into account in your analysis;**
  This is a good point and a discussion on differences in model architectures has been added to the revised manuscript.
- **Why the case studies weren't done for both "finetuned" and "scratch"? How will you evaluate the gain of one against the other?**
  The second sample in the case study has been replaced with a comparison of the first sample to the model trained from scratch in the revised manuscript.
- **Do not miss your goal as this is the scientific question you must answer. You need to structure your analyses so you do not mix up the results;** Yes, we will take this into account in the revised manuscript.
- **2-5, 8: R = 0.5 or R > 0.5? As in Tab. 2. (The same for the other thresholds);**
  This has been clarified in the revised manuscript.
- **Lines 209-210: "*However, a statistical analysis would be wishful to confirm an improving CSI over the prediction time for higher thresholds as shown by the case study in Fig. 8 (a).*" Why haven't you done it yet? You already have the model outputs. This must be included in the results;**
  See our answer to point 9b.
- **Lines 210-211: "*It can be recognized that the HSS, shown in Fig. 8 (b) and (d), provides higher scores than the CSI depicted in Fig. 8 (a) and (c).*" 8, as Figs. 2 and 3, shows the same pattern for CSI and HSS, with different values. You should take a careful look at your results in case you missed something;**
  See our answer to point 8.
- **6:**
  - **Why did you put reflectivity images on the 2nd row? What is the point you want to show?**
    As explained in the caption of Figure 6, the second row is the raw input data. It is used to show what the raw data looks like, compared to our colored versions used for the case studies.
  - **Is it prediction of rain or reflectivity? (See comment 11d.)**

The black and white images represent the rain rate R in mm/h, converted from the raw dBZ RADOLAN data. Clarification and formulas used for this have been added in the revised manuscript.

- **In the color legend, what does it mean when the rain field is gray? The color legend is incomplete;**
  We believe the anonymous referee is referring to the grayscale borders of Germany that were underlaid under the precipitation data to show the 480x480 km cutout of central Germany. We attempt to make this clearer in the revised manuscript.
- **Which forecast times are included in Fig. 6? You should comment this in the caption and in the text;**
  The forecast times used in the case study are explained on page 8, line 170. Clarification to caption and text has been added in the revised manuscript.
- **Where are the "scratch" images?**
  See our answer to point 11d.
- **What does negative reflectivity mean? Why didn't you filter the raw data?** We added more information the data section lines 66 ff, the original data is already gauged: *This data is a reflectivity composite of 17 radar stations in Germany combined with hourly values measured at the precipitation stations. In order to achieve optimized estimates of precipitation, the data on the ground is calibrated with ombrometers. This combination provides high definition data in both temporal and spatial resolution. For more information see Deutscher Wetterdienst (2022)*
-
- **Again, what is the predicted variable, rain or reflectivity?**
  See our answer to point 15b.
- **What is the range of your data so I can understand the differences in the images?**
  Range information for the RADOLAN data has been added in the revised manuscript.

- **Line 220: Here you say "similar", but before you said "slightly better" (line 155);**
  This has been corrected in the revised manuscript.
- **Lines 226-228: "*Comparing the here obtained results with recent publications on deep learning algorithms to precipitation nowcasting based on radar data (Ayzel et al., 2020; Ravuri et al., 2021) the finetuned TrajGRU shows slightly higher scores with less decrease with prediction time.*" See comments 10 and 11;** We added the following sentence in the conclusion: "We notice that Ayzel et al. (2020) evaluate 11 case studies and (Ravuri et al., 2021) consider one case study for their analysis."
- **Lines 228-230: "*While Chen et al. (2020); Ayzel et al. (2020) show that their models are suitable to predict precipitation up to 60 minutes (Chen et al., 2020; Ayzel et al., 2020), we achieve comparable scores for a 100 minute prediction time.*" I couldn't find this analysis throughout the text;** We included the information of the prediction time in the case study section.
- **Lines 230-231: "*A statistical analysis of the prediction time for more than the here presented two case studies would be wishful for future research.*" You have not presented any analysis regarding forecast time (line 209). What you showed here is still not enough for a publication. You already have the data and the outputs of the models, you need to explore further analysis;**
  See our answer to point 9b.
- **Lines 232-233: "To further optimize the results in future studies, different finetuning methods and finetuning hyperparameters could be taken into account." You should add some examples in the text;**
  Examples for methods for possible future studies have been added to the revised manuscript.
- **Line 236: "finally" This is an ambitious comment regarding the precipitation nowcasting problem itself. When completed described, your presented solution can help in some ways, but in my opinion, based on my research and**

experience, I think one product is not enough to solve the complex problem of precipitation nowcasting. I suggest changing to "contribute positively" to be more realistic.
The suggested change has been added in the revised manuscript.

Text comments:

Since most of these points refer to grammar, typographical errors or phrasing, we won't answer on a point by point basis. Most of the suggested changes in this section have been added to the revised manuscript. We thank the anonymous referee again for their thorough comments and suggestions.