EGUsphere, author comment AC2
https://doi.org/10.5194/egusphere-2022-440-AC2, 2022

Reply on RC2

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-AC2, 2022

We would like to thank the anonymous referee #2 for their 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.

- **This study mainly demonstrates the effectiveness of transfer learning with TrajGRU, yet the literature review regarding the transfer learning is very limited and should be much improved with recent studies.** A more extensive literature review has been added in the revised manuscript.
- **In overall manuscript, the explanation of the methodology used by the author, such as model structure, is considered a little insufficient. Although it is specified that the author conducted the research based on the paper of Shi et al. (2017), it would be better to add some more detailed explanation of the research methodology.** A more detailed model structure and methodology explanation has been added in the revised manuscript.
- **In abstract Line 5, the authors clarified that *"In this work a Convolutional Long Short-Term Memory network (ConvLSTM) is applied to Radar data of the German Weather Service."* Although they used TrajGRU in this study and mentioned the comparison between ConvLSTM and TrajGRU in Section 2.2, I don't understand why the authors said "ConvLSTM" instead of "TrajGRU".** This is indeed a mistake, it should say TrajGRU. This has been revised.
- **In section 2.2., the authors described the main formulas of TrajGRU, some notations of the equations are missed (e.g., □, f). In addition, there seems to be a lack of explanation for the comparison between ConvLSTM and TrajGRU, especially for figure 1. Also, please add more information in caption of figure 1 (e.g., what is colored lines mean?).** Missing notations and missing captions for figure 1 have been added in the revised manuscript.
- **In Line 118, *"Because of this we freeze the weights of the outermost TrajGRU layer of both encoder and forecaster for the finetuned model and only train the two innermost layers on the German RADOLAN data afterwards."*, please provide additional information about how you fine-tuned model (e.g., learning rate setting, etc.). Furthermore, I wonder if the authors experimented with directly using pre-trained parameters in the new model on RADOLAN data. Although they mentioned that *"Other finetuning configurations were tested, such as freezing more layers or none at all, but displayed worse performance."* I suggest adding more detailed explanation of the other possible fine-tuning approaches used.** A detailed experimental setup of the

finetuned model (for example learning rate) can be found on Page 5 Line 110 to Line 114. A more detailed breakdown of the other finetuning configurations was added in the revised manuscript.

- **The general purpose of transfer learning (i.e., fine-tuning) is to solve the problem of model underfitting due to the limited availability of model input data. The authors also explained this, and information on the amount of RADOLAN data used is given, but there is a lack of information about the amount of pre-trained HKO-7 data. Therefore, it would be better to discuss not only the distribution of data according to rainfall intensity but also the difference in the overall amount of data between RADOLAN and HKO-7 data.** A more detailed discussion of the differences in overall amount between RADOLAN and HKO-7 has been added in the revised manuscript.
- **Since the model performance fluctuates with increasing the number of iterations, and there is a possibility that overfitting problems should seriously affect the model performance, it is difficult to say that the fully trained model result (i.e., performed initial set 100,000 iterations) is the optimal result. So, I wonder if the author used any method other than full training to reach the optimal model state used to obtain the highest scores and the number of iterations it takes to reach it, as mentioned in Tables 2 and 3. If not, it is suggested to use a methodology such as early stopping method to obtain optimal model performance.** The results in the paper are obtained using the model iteration that had the best average CSI and HSS scores. Some explanation about this process as well as general thoughts about more optimal methods like early stopping or weight decay have been added in the revised manuscript.

- **For the results of case studies, in figure 6, it has not been fully explained why two different pictures of the input data are needed. Are there any implications for each of the two pictures? If not, it would be better to remove one of the two. I suggest from the perspective of comparing model results, the picture in the second column is better to remove. Also, why don't you compare the "train from the scratch" model results with "finetuned" model results in case studies? It would be interesting to see the effect of fine-tuning through qualitative comparison with the "train from scratch" model.** We thank the anonymous referee for the suggestion. We agree that the second sample is redundant and a direct comparison with the model trained from scratch would be more interesting, so we replaced the second case study with a comparison to the model trained from scratch in the revised manuscript.