Interactive Discussion: Author Response to Referee #1

Brief Communication: Training of AI-based nowcasting models for rainfall early warning should take into account user requirements

Georgy Ayzel and Maik Heistermann
EGUSphere, doi:10.5194/egusphere-2024-1945

RC: ***Reviewer Comment***, AR: *Author Response*, ☐ Manuscript text

Dear referee,

thank you very much for your positive response, and for the time and effort spent to examine the manuscript.

The comments are very useful and will be comprehensively considered in the revised version of the manuscript. Please find a point-by-point reply below.

Thanks again for your willingness to review this manuscript.

Kind regards,
Maik Heistermann
(on behalf of both authors)

RC: ***[...] As recognized by the authors, the scientific significance of the result is rather limited, because one would obviously expect the scores of a statistical prediction method to benefit from training with forecast ranges, intensities and weather regimes consistent with those used as verification. Nevertheless, the paper has some educational value for pointing out to the machine learning community that statistical weather prediction models optimized with respect to generic metrics cannot, in general, be expected to perform competitively in terms of more specific metrics.***

AR: We agree that one would "*expect* the scores of a statistical prediction method to benefit from training with forecast ranges, intensities and weather regimes consistent with those used as verification". Still, it is, in our opinion, neither obvious nor self-evident that this expectation would actually materialize, since all models are verified on the same data. A possible outcome could as well have been that, even if trained to predict a specific threshold, the corresponding model could *not* outperform one that was trained on a more generic metric. In fact, this is what we observe for increasingly high thresholds: as predictability deteriorates, all models essentially fail (most evident for the 40 mm/h threshold).

We had pointed out this finding in ll. 154-156 of the preprint:

> These results are in line with our hypothesis that making the training task more specific pays off by a higher predictive skill. One might argue that this result is unsurprising. In our view, though, it is by no means self-evident that the segmentation models could actually capitalize on a more specific training task.

Maybe it would be worth to revisit this statement once more in the conclusions section.

Despite these findings, we entirely agree with the referee that the main point of this paper is of "educational value". This is why we reiterate in ll. 190-192 of the conclusions that

> [...] the aim of our study was not to introduce superior DL architectures or model structures, but to demonstrate how a simplification of the training task can help to improve model skill and to boost the usefulness for specific user groups.

This is exactly why we chose the format of a "brief communication" over a "research paper", with all the consequences e.g. in terms of technical detail of model architectures (see comment below).

RC: ***The paper has 11 pages, which seems long for brief communication according to the NHESS website (it indicates a limit of 4 pages).***

AR: As far as we know, the recommended limit of 4 pages for the brief communication format refers to journal pages. As compared to the preprint, the final typeset manuscript will be substantially shorter. Based on our previous experience with this manuscript type (and also considering other brief communications in NHESS), a length of 8 preprint pages (up to the end of the conclusions) plus references corresponds well to the required page limit.

RC: ***Not enough information is provided to understand key technical aspects of the study: in particular, there should be a more detailed description of the original RainNet architecture and of the motivation for its changes : what is EfficientNetB4? What is the relevance of LogCosh loss to heavy precipitation? How do you define the threshold on the output to obtain 'a segmentation task'?***

AR: The lower extent of technical detail is owed to the required brevity of the present manuscript type, together with the fact that the main point of the manuscript does not follow from the technical details (see first comment).

Still, we understand that, for parts of the audience, a more detailed technical or methodological description might be desirable. Originally, we had hoped that providing the code repository including its documentation would provide the required technical details to interested readers; however, we agree that an intermediate level of methodological description would be helpful. At the same time, the length of the manuscript is already at the upper limit, and we are convinced that adding such detail in the manuscript would not be helpful to bring across our key message. We therefore suggest to add a supplement to the paper in which we describe in some more detail the abovementioned aspects of e.g. model architectures or loss functions, and refer to this supplementary from the main manuscript.

RC: ***False alarms are a key problem when issuing warnings. The dataset used was taken from CatRaRE catalog, which is designed to contain observed extreme events. It means that the training and testing datasets are biased. Thus, the verification will likely underestimate the false alarm frequency, because the dataset excludes cases where the models generate a heavy precipitation forecast and none was observed. A solution would be to compute scores over the entire 2019-2020 testing period. Alternatively the paper should present some proof that the data sampling does not affect false alarm counts. This is the main scientific issue of the paper.***

AR: This is a valid concern. However, even when we focus on situations with heavy rainfall, the model domain of 256 km x 256 km is typically dominated by rainfall accumulations below our thresholds, i.e. by "non-event" grid cells. For our testing data, the frequency of such "non-event" grid cells amounts to 95.73 % for the

threshold of 5 mm in one hour and increases further with increasing precipitation thresholds (10 mm: 98.84 %, 15 mm: 99.56 %, 20 mm: 99.81 %, 25 mm: 99.91 %, 30 mm: 99.96 %, 40 mm: 99.99 %). So, even though the data samples obviously contain enough threshold exceedances for the models to learn, the dominance of "non-event" grid cells prevents overprediction (please also see our response to the referee's below comment), and we would not consider the data sets biased from an application perspective.

In order to address this comment, we will briefly discuss this issue in the manuscript, and add a table in a supplement that provides, for each investigated threshold, the frequencies of events (threshold exceedances) and non-events in the testing data samples.

RC: ***Significance testing is missing from the results, which is problematic for a paper about statistical prediction. At least, figures 2 and 3 should display some confidence intervals.***

AR: We agree that a measure of significance would make sense in Fig. 2, particularly for very high accumulations for which the skill of all models becomes very low. We will add confidence intervals for the CSI based on a resampling/bootstrapping procedure. For Fig. 3, however, we would prefer to keep the conventional FSS presentation format (please also see our below response to the referee's suggestions on Fig. 3).

RC: ***The PySteps system is getting old. It would be more convincing to display the scores from a more recent nowcasting system such as DGMR, as a performance baseline.***

AR: Generally, we would agree it would be interesting to include other recent DL-based nowcasting schemes in this analysis. However, the main motivation of the study is not to provide a benchmark experiment among architectures, but to highlight the relevance of considering user requirements in the training task. As we pointed out in the conclusions (ll. 192-193 of the preprint), "this approach should be systematically explored also for recently proposed DL models" (including e.g. DGMR or NowcastNet). Besides, PySteps is still considered a competitive benchmark model representing the conventional approach of optical flow and Lagrangian persistence, and we consider it good practice to include it in the present study. We would therefore prefer to keep the model selection for the benchmark experiment as it is.

RC: ***The mention of 'user requirements' in the title sounds a bit excessive, because generating rainfall warnings involves other considerations than the choice of accumulation period and threshold. It may be more appropriate to state that the paper demonstrates the sensitivity of nowcast performance to the choice of objective function.***

AR: We understand the referee's reservations with regard to the title. In essence, any objective function is (or should be!) a representation of what a user expects the model to predict. Certainly, user requirements can go beyond the objective function, e.g. in terms of computational efficiency, transparency, comprehensibility and more. However, as we repeatedly addressed in this response, our manuscript aims at reducing the level of technical detail. Our hope is that for the non-technical parts of the audience (as might be the case in NHESS in comparison to other journals), our current title might be more informative, as compared to explicitely referring to objective of loss functions. Besides, we more generally (or unspecifically, if you may) refer to "training" in the title, instead of "objective function", as the training also involves the entire setup including the selection of training data. Of course, the ability of the title to comprehensively reflect the content of the paper is also limited, which is why we explicitely and specifically explain how we define potential user requirements in the context of our study, and make clear that this is only an example.

Altogether, we hope we could explain our motivation of using the term "user requirements" in the title, and we would be glad to keep the title as is due to the given reasons.

RC: ***Please clarify how can the Jaccard loss can be differentiated, since the Jaccard metric is a ratio of integer***

success counts.

AR: The Jaccard loss is a relaxed, differentiable modification of the Jaccard index, which is a count-based measure similar to the Intersection over Union (IoU) metric and the Critical Success Index (CSI). It describes the ratio between hits and the sum of hits, misses, and false alarms. To address the issue of differentiability, Rahman and Wang (2016) proposed the following relaxation of the Jaccard index:

$$I_{Jaccard} = \frac{\sum_i y_i \hat{y}_i}{\sum_i y_i + \sum_i \hat{y}_i - \sum_i y_i \hat{y}_i}, \tag{1}$$

where:

- y_i is the ground truth binary label for the i-th pixel or element.
- $\hat{y}_i$ is the predicted probability for the i-th pixel or element.
- $\sum_i y_i$ is the sum of the ground truth binary labels (the count of positive ground truth samples).
- $\sum_i \hat{y}_i$ is the sum of the predicted probabilities (the count of positive predictions).
- $\sum_i y_i \hat{y}_i$ is the sum of the element-wise multiplication of the ground truth and the prediction (the count of true positive samples).

Hence, the Jaccard loss function becomes

$$\mathcal{L}_{\text{Jaccard}} = 1 - I_{Jaccard}. \tag{2}$$

In this way, the gradient of the Jaccard loss function can be computed and integrated into the optimization routine for finding parameters of the neural network.

In the revised version of the paper, we will add the reference to Rahman and Wang (2016).

RC: ***The CSI score should be complemented by some information about the hit rate and false alarm rates (false negatives and false positives), as both are very important for the credibility of warnings.***

AR: The CSI is already quite a balanced score as it takes into account hits, false alarms and missed events. Given that (even in our samples which present a focus on heavy rainfall situations) the "non-event" grid cells are still dominant, the CSI cannot grow much at the cost of increasing false alarms (see our response to your above comment).

Still, we are willing to compute both hit rates (probability of detection, POD) and false alarm rates (FAR). In order to keep the manuscript brief and follow the requirements of a brief communication, the corresponding figures/tables will be provided in a supplement. We would like to ask for your understanding that we do not already provide the FAR/POD numbers in this interactive discussion. The reason behind this is that we first need to recompute the entire verification since predictions were not stored due to the resulting massive data volumes.

RC: ***typo on line 190 'prediciton'***

AR: Thanks for spotting the typo which will be fixed in the revised version.

RC: ***Figure 3 is hard to read in terms of comparison between the systems. Since there is not much information in the dependency on scale, it may be better to present curves of FSS(range) at a fixed scale (say, 20km), instead. It would also facilitate the display of confidence intervals or statistical significance of the FSS differences.***

AR: We do not agree that there is "not much information in the dependency on scale". In our view, demonstrating the dependency on scale is the main motivation of this figure (see also the comments of referee 2). Of course it is obvious that the FSS will increase (or at least not decrease) with increasing scale. The amount of increase is, however, not obvious beforehand, and apparently differs both between models and precipitation thresholds. In our opinion, this figure is not so much about distinguishing exact differences between individual cells, but rather about giving an intuitive and easy to grasp representation of how the FSS varies with model, spatial resolution and precipitation threshold. The actual values printed in the cells are merely a support for those interested in a closer inspection; yet, we do not consider confidence intervals as helpful in the context of this figure (and also not common in this kind of FSS diagrams).

However, if we print the FSS values in the cells, we agree that the legibility could be improved. To this end, we suggest to rearrange the three panels by putting them on top of each other (i.e. 3 x 1 matrix) instead of beside each other (1 x 3 matrix). That way, the plot could be larger, so that the font size of the labels can be increased, too.

References

Rahman, M. A. and Wang, Y.: Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation, in: Advances in Visual Computing, edited by Bebis, G., Boyle, R., Parvin, B., Koracin, D., Porikli, F., Skaff, S., Entezari, A., Min, J., Iwai, D., Sadagic, A., Scheidegger, C., and Isenberg, T., pp. 234–244, Springer International Publishing, Cham, ISBN 978-3-319-50835-1, 2016.