EGUsphere, referee comment RC1
https://doi.org/10.5194/egusphere-2022-1159-RC1, 2022


# Comment on egusphere-2022-1159

Anonymous Referee #1

Referee comment on "Shallow and deep learning of extreme rainfall events from convective atmospheres" by Gerd Bürger and Maik Heistermann, EGUsphere, https://doi.org/10.5194/egusphere-2022-1159-RC1, 2022

In their manuscript, Bürger and Heistermann trained and applied multiple ML/DL models to the (binary) classification task of detecting convectively enforced extreme rainfall in Germany. They compare a whole range of different models ranging from regression models through random forests and shallow neural nets to well-known DL models for image classification like AlexNet, GoogLeNet or ResNet. The study's classification task can be formulated as follows: Given the (daily aggregated) ERA5 fields of CAPE, convective rainfall, and total column water over Germany, does CatRaRE contain at least one event exceeding warning level 3 where the event does not exceed nine hours. Bürger and Heistermann used cross-entropy as the loss function during training and evaluated their final results based on the Equitable Threat Score. As their DL models use a stochastic gradient descent algorithm during the optimisation, they trained 20 individual realisations per DL model resulting in an ensemble. Based on that (simplistic) ensemble, they report that the ALL-CNN model shows the highest mean ETS (across the ensemble) of .52. At the same time, the ResNet architecture provides the ensemble member with the highest individual ETS (.54).
The manuscript is well structured, and I appreciate the extensive model selection used for comparison and acknowledge the effort spent to train all of these. Even though the intercomparison of different methods and architectures is interesting on its own, I have difficulties distilling the overall relevance (concrete use case) of the classification for meteorological applications. I have some more points of concern, as listed below, but I am confident that the authors can adequately address them and that the manuscript can provide an important contribution to the field. I wonder if the GMD/ESSD inter-Journal SI "Benchmark datasets and machine learning algorithms for Earth system science data" (https://gmd.copernicus.org/articles/special_issue386_1147.html) might be suited better to the manuscript's scope. At least for me, the study's focus lies more on the intercomparison aspect rather than the "monitoring of precursors of evolution".

## Major Comments

- As mentioned above, it does not become clear to me what consequences a statement like "There is an extreme convective event (somewhere) over Germany" might have for a meteorologist, climatologist or decision-maker. L 229f somehow reflects the ultimate goal; however, it might be good to further distil the gain also in the introduction.
- I wonder how a cross-entropy or ETS analysis might contribute to a better understanding of the influence of 'deep' in DL models, as stated in l. 41f. For such a statement, I would have expected some explainable AI (XAI) methods or some sensitivity analysis of each model type, like varying the number of inception blocks in the 'GoogLeNet-style' model. Here the introduction raises expectations that the conclusion does not reflect.
- As far as I understand, you are using ERA5 data (cape, cp, tcw) as input **X** and CatRaRE as target **y** for training (2001-2010) and validation (2011-2020). Finally, you apply the trained model to data from HIST and RCP85. In l 144, you correctly state that the second dataset is not independent of the DL models, as you use those for model selection. As overfitting can happen on both - parameters (training set) and hyperparameters (validation set), why do you not split your data into three sets (training, validation, test)? Especially as you apply the trained models to data from different sources that likely have different properties, I think it would be beneficial to compare the test set's performance against the same (sub-)period of RCP85. Thus, you could detect differences in model performance that might serve as a guide towards interpreting all RCP85 data where you do not have any labels.
- I suggest broadening the analysis of the predicted probabilities over the entire detection period. For example, replacing Fig. 5 with a reliability diagram where the predicted probability is plotted against the observed relative frequency might reveal model-specific differences.
- Given the close range of ETS values across the different models, I suggest providing uncertainty quantifications and/or statistical tests to demonstrate the significance of your findings.
- How do already existing 'classical' findings of the expected change of extreme precipitation align with your classification results? Can you discuss the concept drift in the data that the classifier faces?
  - In that regard, which period do you use to calculate the mean and std for the z-transformation?

**Minor Comments**

- L. 22ff Besides the references to the 'classical' DL introductions, I encourage the authors to also focus on the recent discussions on ML/DL applications in atmospheric sciences like Reichstein et al. (2019) and Schultz et al. (2021).
- L. 158f How do you analyse the influence of cape? In l. 126 you state that you are using cape, cp and tcw as channels similar to RGB. Please clarify how you create the "non-cape" classifications. Do you train the models with two channels only? Do you replace the cape channel with zeros or another variable?
- Fig. 1 shows cape values jointly with the CatRaRe events used to define the extreme labels. The selected model domain contains pixels outside of Germany. CatRaRE, however, covers Germany only. Did you check (most likely with some other dataset) how often (if at all) extreme events occur outside of Germany but within your defined model domain? For me, that seems to be a potential source of introducing labelling errors.

- Fig. 4: I suggest using a more colourblind-friendly palette.
- Even though Table S1 lists several tuned hyperparameters, how does the learning rate change under the poly policy?
  - I suggest adding a column reporting the number of trainable parameters of your modified versions

- Did you consider also using architectures already focussing on precipitation (for example (your) RainNet model (Ayzel et al., 2020)) and adjusting details for your classification task?
- L. 59 I am wondering if a log transformation for cp before applying the standardisation might be beneficial
- Please provide some more details on the EOF reduction. For example, how many components are you using?
- From the first sentence in your abstract, I expect this manuscript to focus on creating a new data set that can be used for ML/DL applications. In its current state, the abstract does not adequately transport the enormous (DL-)model comparison you performed.

**Formal Comments**

As a reviewer, I was asked helping to ensure that manuscripts comply with the journal's guidelines. Therefore I'd like to point out some formal aspects:

- Please add a "competing interests" statement as required by Copernicus Publication (see https://www.natural-hazards-and-earth-system-sciences.net/submission.html#manuscriptcomposition §16)
- Software Code: You refer to your GitHub repository but to the best of my knowledge Copernicus Journals prefer software provided through a DOI (e.g. through zenodo)
- URLs: Please add the last access dates to all URLs
- A legend is missing in Fig. 3

**References**


- Ayzel, G., Scheffer, T., and Heistermann, M.: RainNet v1.0: a convolutional neural network for radar-based precipitation nowcasting, Geoscientific Model Development, 13, 2631–2644, https://doi.org/10.5194/gmd-13-2631-2020, number: 6, 2020.
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.
- Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S.: Can deep learning beat numerical weather prediction?, Philosophical Transactions of the Royal Society A: Mathematical, Physical and

Engineering Sciences, 379, 20200 097, https://doi.org/10.1098/rsta.2020.0097, 2021.