Here we substantiate our first response given to reviewers A and B at https://egusphere.copernicus.org/preprints/2022/egusphere-2022-1159. Line numbers refer to the new text.

We note that due to an error in the ETS estimate (the probability threshold for validation was optimized and not taken from the calibration set), the corresponding results have shifted so that now TREE has the best ETS score.

Rev #1

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.

Please see the new § (l. 62-73) which explains more clearly our intention to "to raise awareness among researchers and decision makers for an impending change in these statistics".

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.

See comment above and also l. 279-287.

- 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.

We give our interpretation of the results depending on network architecture in the new § at l. 194-212. This includes a discussion on width and depth of a network and if extended complexity is necessary in our case or not.

- 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.

This topic was already discussed in the discussion phase, where we argue that overfitting is not to be expected; see also §2.4 of the text.

- 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.

We have added reliability and sharpness in Figs. 5 and 6, plus the new §3.3.

- 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.

Stochastic uncertainty in shallow modeling was missing in the discussion paper. We have now reformulated TREE and NNET as ensemble models and consider, like in the DL case, corresponding means. The SI presents a second ('cloned') realization of all models demonstrating that the results are essentially stable with respect to the main conclusions, l. 189-193.

- 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?

We have expanded on the "common wisdom" as reported by the cited IPCC source, see l. 291-301. For the concept drift we have a new § at l. 266-278.

- In that regard, which period do you use to calculate the mean and std for the z-transformation?

It is 2001–2020, as mentioned in l. 91.

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).

References are added.

- 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?

We added a corresponding footnote at l. 178.

- 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.

Here we disagree, the German border has no relevance for the classification. If a pixel falls outside, it may only mean that it is too distant to affect the local event, a fact that should be learned by the schemes.

- Fig. 4: I suggest using a more colourblind-friendly palette.

We have put the updated Fig. 4 through Coblis and it looks good there.

- Even though Table S1 lists several tuned hyperparameters, how does the learning rate change under the poly policy?

The hyperparameters are described in the SI. The learning rate follows a polynomial decay, $(1 - \text{iter}/\text{max\_iter})^{\text{power}}$, becoming zero when max_iter is reached. This is added as a footnote on l. 137.

- I suggest adding a column reporting the number of trainable parameters of your modified versions

We have added a corresponding column to Table 2 (which describes the network architecture).

- 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?

No. RainNet is used to map one state of a system to another of the same system, whereas here we need to map one state of one system to another of another system. Or, in other words, RainNet was designed to capture the motion and intensity dynamics of precipitation fields at very high spatial and temporal resolution, and required much larger amounts of data for training (several years of five minute data). We would not consider RainNet as specifically suitable just because it aims at the same variable, precipitation. The processes or relationships learned by RainNet are very different from our setup. We do not say that RainNet is unsuited for the task, but our approach of selecting candidate DL models was a different one (using established models for image classification).

- L. 59 I am wondering if a log transformation for cp before applying the standardisation might be beneficial

We discuss this in a footnote on l. 87.

- Please provide some more details on the EOF reduction. For example, how many components are you using?

This was added at l. 117.

- 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.

We changed the abstract accordingly.

Formal Comments

- Please add a "competing interests" statement as required by Copernicus Publication (see https://www.natural-hazards-and-earth-system-sciences.net/submission.html#manuscript-composition §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

All have been addressed, except: Our repository is gitlab, to which unfortunately no connector from Zenodo exists; there are no urls in the References; legend info added in the caption.

Rev #2

We would like to thank the referee for taking the time to review our manuscript, and for the open and clear criticism. The central part of the comment is, as we understand, as follows:

„[...] [the manuscript] basically [...] is not driven by any research questions and objectives. Without finely tuning each method, especially for the sophisticated CNN models, it is unclear why to compare these methods and also the very similar results for each method give readers very limited insights from their studies.“

We have revised the manuscript to better convey the research questions of our study. Please see the tracked changes, in particular sections 1, 3.2, 3.4 and 4.

The results pertaining to the DL models are inconclusive because their application requires more fine-tuning.

We repeat that with regard to DL tuning, we have described that we purposefully used the basic model structure *as is* from the corresponding image recognition tasks, but fine-tuned settings to achieve convergence in the learning curves. Model uncertainty, moreover, is a genuine part of DL and thoroughly covered by Brownlee (2018), we have addressed it in greater detail in the abstract, §3.2 and §4.

The results are not relevant to the reader because the performance of all benchmarked models is very similar.

While we do not understand the argument (as addressed in the paper discussion), after fixing the above-mentioned error the results are no longer very similar (and have not been before).