We thank the reviewer for her/his insightful comments, which demonstrate a thorough understanding of the topic, including the weak points of this study to which we respond here. Reviewer comments are quoted in red.

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.

The relevance lies entirely in the application to climate projections, that is, inferring future information about CatRaRE-type events. The “concrete use case” is therefore not a warning like “tomorrow there will be a CatRaRE event somewhere in Germany”, but a corresponding one that for future summers this type of event will become more frequent. We will add a clarifying statement in the introduction.

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.

Yes, and it indeed remains to be shown whether or not a Germany-trend applies more or less uniformly to any of Germany’s sub-regions; it is what we aim at in a geographically refined follow-up study. Our focus here was on methods, and since we had so many of them the danger was that regional detail would confound the results and make the text less readable.

- 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.

The ’deep’ in DL was meant to assess the added value of using the novel deep methods as compared to the conventional (”shallow”) ones; and added value is meant here solely in terms of classification skill. But we agree that the conclusions don’t hold what was foreboded earlier, so we will add clarification accordingly around Table 3.

- As far as I understand, you are using ERA5 data (cape, cp, tcw) as input $\mathbf{X}$ and CatRaRE as target $\mathbf{y}$ 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.

On l. 144 we meant the validation dataset, i. e. ERA5 and CatRaRE from the period 2011-2020. These data and the estimated empirical models are, in a statistical sense, not fully independent. The corresponding hyperparameters (cf. Table S1) were selected based on the convergence of the learning curve, and we are confident that for new test data learning will converge similarly; classification skill was not a selection criterion. It is true that a triple-split into training, validation, and testing would have been the more stringent approach. We feared, however, that not much would be gained from that due to the reduced dataset length and corresponding added uncertainty.

- 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.

That is indeed a useful idea, and we will add both reliability and sharpness diagrams. The latter reveal that most DL predictions are quite sharp (close to binary predictions), more sharp than the shallow models.

- 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.

Because the sampling uncertainty of ETS (and most likely of all other skill scores) depends not only on the contingency table but also on the underlying data distribution, an idea about ETS uncertainty can heuristically be obtained by slightly perturbing selected experiments (e.g. using slightly different CatRaRE thresholds) and check their outcome. This will be added.

- 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 will expand on the ”common wisdom” as reported by the cited IPCC source.

- 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. 64/5.

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).

We welcome the mentioning of these two instructive paper which we shall include in the references.

- 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 trained with the remaining predictors, and will clarify the text accordingly.

- 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 tried hard to render Fig. 4 using a colorblind-friendlier palette. At this stage, we decided against it because none of the rendering was able to sufficiently separate the different colors (4 for Shallow, 9 for Deep) across all possible color vision deficiencies. Unless required by the publisher we prefer to keep the current version.

- 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 will be added in the SI.

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

We will add 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

That is an interesting option. We were (naively) confident that the log-based network transfer functions take care of that, but maybe not; and the Shallow models do not anyway. We will discuss this, and if beneficial we may even apply it for the revision.

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

This will be added in the revision.

- 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.

It will be changed.

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#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

All will be addressed.