The paper by Bürger and Heistermann applies a variety of machine learning algorithms to the task of classifying atmospheric fields as conducive to severe convection in Germany. Trained and evaluated on ERA5 and CatRaRE, the models are applied to EURO-CORDEX simulations for the past and future to analyze trends in the potential for severe convection.

This paper stands out from the mass of applied machine learning papers by actually comparing a full range of simple, intermediate and very complex models. Given the great recent interest in such methods, I think that this kind of study can be valuable to the natural hazards community, as well as climate science in general. The paper is overall sufficiently well written and the methodology seems sound to me. I nevertheless have a number of concerns with the current state of the manuscript, which require some further revision before I recommend its publication.

General comments:

- The motivation for this work should be clarified in the introduction: Make it clear that it is far from obvious that the "killer apps" from image classification contests will be similarly worth their cost in a weather and climate context. I see a number of reasons for doubting their applicability:
    - atmospheric fields have very different spatial statistics from images of cats and dogs
    - your target quantity (severe convection or not) exists on a spectrum. As far as I'm aware, there are no animals that are "close to the threshold" of being a cat or dog.
    - the amount of training data is often severely limited in our field
    - unlike MNIST and the like, your "images" might have long term trends (see also my comment below)

It is therefore a good idea to actually test whether there is any benefit in the "deeper" approaches over classical machine learning / statistics.

- The conclusions should be made more clear as well. I would argue that you found no substantial benefit of deep over shallow methods (NNET for example seems competitive with its deeper sibling models but is so much simpler). That is an interesting result and arguably good news for researchers with limited expertise in deep learning who can thus rely on relatively cheap and simple methods.
- To what extent do you think that your conclusions can be generalized to other similar studies? I guess one thing to keep in mind is that hazards are usually rare whereas you define your problem in a way that leads to a balanced dataset.
- I have several comments on the issue of trends in your data:
    - It is well known that neural networks are typically bad at extrapolating

outside of the training range. This could play a role here, depending on how exactly you normalized the data from ERA5 and the CORDEX simulations. Did you estimate the mean and standard deviation from the whole period? Or just the training and / or validation time?

  - Are the trends consistent across your 20 optimization runs or do they randomly differ?
  - Did you look at the trends in your predictor variables? If ERA5 and CORDEX have different trends in, for example, CAPE, couldn't that easily explain the discrepancy between HIST and ERA5 trends?
  - For trends in ERA5 convection parameters, you can also refer to https://doi.org/10.1038/s41612-021-00190-x

- I think that "explainable AI" methods like Shapley values (which are model agnostic, see https://doi.org/10.48550/arXiv.1705.07874) would have been helpful in actually understanding the differences between the different models. This could have explained which model uses which variables and which of the variables are responsible for the trends. It is likely too late to add such analyses to the manuscript, but it could be mentioned in the conclusion / outlook section and perhaps taken into account in future studies.
- I am a little concerned about the use of convective precipitation (cp) in this study. The amount and spatial distribution of rainfall produced by the convection parametrization depends heavily on (a) the type and settings of the parametrization and (b) the internal horizontal resolution of the model. The resolution definitely differs between ERA5 and COSMO. I'm not sure how similar the parametrizations are, I guess they are both based on the Tiedtke scheme? In light of possible fundamental differences, how can you apply a model trained on ERA5 to cp fields from another model? Wouldn't it be wiser to use total precipitation instead, which is more generally comparable?
- Is there a reason why no dynamical variables like wind shear were included?

Specific comments:

l.18 and other places: please explain somewhere, what exactly you mean by "saturation effect".

l.62 "the best performing models are applied" I believe you apply all of them, which is a good idea.

l.92 "RCP85 (2006-2020)" didn't you use the full scenario run from 2006-2099? Or does this sentence only apply to the climatology used as reference? Either way please clarify over which part of the time series you estimated the mean and standard deviation for your normalization in each case. In particular when applying your models to the full ERA5 time series, did you use the same normalization as in training or normalize over that whole period? I think this could make a difference for the resulting trends.

l.114: “we employ four shallow statistical models:” but then you list only three, forgetting NLS

l.116 please explain what exactly you mean by EOF orthogonalization. Did you apply eofs to the fields and use the principal components as predictors? If so, which part of which time series were the EOFs estimated from?

l.116 How did you arrive at the seemingly random number of “33, 27 and 21” EOFs?

l.124 How is “Logreg” a deep method with only one layer? And how does it differ from your logistic regression (NLS)?

l.126 please upload the code to some permanent repository like zenodo, as per journal policy. I don’t know what you mean by “no connector from Zenodo exists”, why can’t you just upload your code there?

l.152 what do you mean by “EOF truncation”? The same as EOF orthogonalization above?

Fig.2 and 3: why is the test loss so much smoother than the training loss?

l.175 what do you mean by “optimal probability threshold”? Don’t you just predict a class if its probability is greater than 0.5?

Fig.4 please explain why you are specifically interested in the importance of cape? what conclusions do you draw from this?

l.268 “reasons unknown” is it so surprising that convective activity would have different trends in ERA5 and COSMO?

Table 3: Is it correct that none of the HIST trends are significant (not boldface) despite their relatively large magnitudes?