https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

Deep learning of extreme rainfall events from convective atmospheres

Gerd Bürger[1], Maik Heistermann[1]

[1]Institute of Earth and Environmental Science, University of Potsdam, Potsdam, 14476, Germany

Correspondence to: Gerd Bürger (gbuerger@uni-potsdam.de)

Abstract. Our subject is a new Catalogue of radar-based heavy Rainfall Events (CatRaRE) over Germany, and how it relates to the concurrent atmospheric circulation. We classify daily atmospheric ERA5 fields of convective indices according to CatRaRE, using an array of conventional statistical and more recent machine learning (ML) algorithms, and apply them to corresponding fields of simulated present and future atmospheres from the CORDEX project. Due to the stochastic nature of ML optimization there is some spread in the results. The ALL-CNN network performs best on average, with several learning runs exceeding an Equitable Threat Score (ETS) of *0.52*; the single best result was from ResNet with *ETS = 0.54*. The best performing classical scheme was a Random Forest with *ETS = 0.51*. Regardless of the method, increasing trends are predicted for the probability of CatRaRE-type events, from ERA5 as well as from the CORDEX fields.

1 Introduction

Since computing power has grown to levels that were beyond imagination just years ago, automated and numerically expensive (machine) learning has evolved into a versatile and capable tool set for data science. This applies in particular to *Deep Learning* (DL), which refers to neural networks with a notably increased number of neuron layers. Many scientists are now curious whether their older, conventional models can stand the test of skill against these newer methods. Examples are abundant, for example from climate simulations and weather prediction (daily to seasonal) (Gentine et al., 2018; Ham et al., 2021, 2019; O'Gorman and Dwyer, 2018; Rasp et al., 2018; Weyn et al., 2021). Generally, DL is evolving with such a speed that makes it hard to keep pace; for a general introduction into Deep Learning, (Bianco et al., 2018; Goodfellow et al., 2016; Alzubaidi et al., 2021) provide a nice and thorough overview. At least in the data driven disciplines, hence, one may be in hope or in fear about the perspective that much of the scientific progress of the past several decades is about to be dwarfed by machine learning techniques.

In this study we aim to explore the potential of DL in the field of atmospheric weather types (classification). We investigate synchronous daily sequences of large- and local-scale weather patterns over Germany. As predictors we use reanalyzed atmospheric fields whose spatial resolution is coarse enough to permit long climate model projections. These fields are 'labeled' by the occurrence of local, impact-relevant extreme convective rainfall events anywhere in the study

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

area. The events were obtained from a recently published catalogue of extreme precipitation events in Germany (CatRaRE, (Lengfeld et al., 2021)) which in turn is based on a 20-years record of gridded hourly radar-based precipitation estimates (RADKLIM, (Winterrath et al., 2018)).

By interpreting each atmospheric field as the color code of a 2-dimensional "image", our task can be framed as one of image classification. Given the geometry and resolution of the fields (cf. section 2), the classification is done in a space of dimension ~4k. This number roughly compares to some of the classical DL datasets such as MNIST (dim. ~1k) and CIFAR-10 (dim. ~3k), but is certainly small compared to newer sets such as ImageNet (dim. ~100k) or Open Images (dim. ~5M), cf. Table 2. Likewise, while most of the DL networks have to choose between as many as 1000 classes, our initial example is just binary. Therefore, if CatRaRE-relevant patterns of atmospheric moisture over Germany can be compared at all to images of cats and dogs, one could naively expect a performance that is comparable to published classification results on those image datasets.

Our focus shall generally not be on obtaining the best result currently possible, but instead of better understanding the influence of the 'deep' in DL. To that effect, we have explored a number of conventional and newer 'shallow' methods, and compare them to a selection of DL networks that, each in its time, had entered the DL arena quite spectacularly; an overview of the used methods is given in the Supplemental Information (SI). Our DL framework is Caffe, which provides a genuine Octave/Matlab interface to DL (Jia et al., 2014). The Caffe framework along with most of the networks have already seen the height of their days, and are by now being superseded by more sophisticated and successful networks and frameworks (Alzubaidi et al., 2021). This only indicates that the development continues to be fast, making it difficult to keep pace. By not trying to keep pace, our focus lies on the historical context and on an understanding of the effects of 'Depth' on the performance.

After analyzing the performance of the various methods and exploring the difference between shallow and deep approaches, the best scoring methods are applied to simulated atmospheres from the EURO-CORDEX project (Jacob et al., 2020); the predicted classification can be used to estimate past and future changes in the frequency of extreme precipitation events of the type contained in the CatRaRE catalog.

2 Methods and Data

2.1 Atmospheric data

Since our focus is on convective events, we restrict the analysis to the warmer months from May to August. From the ERA5 reanalyses (Hersbach et al., 2020), atmospheric convectivity is measured by the indices of convective available potential energy (cape), convective rainfall (cp), and total column water (tcw). They are used as potential classificators, given as daily averages over the area between the edges [5.75E 47.25N] and [15.25E 55.25N], normalized with, for each variable, mean and standard deviation across time and space. Future atmospheric fields are obtained from the EURO-

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

CORDEX initiative and are simulated by the model CNRM-CM5 (simply "GCM" in this text) driving the regional model COSMO-crCLIM ("RCM"). We use emissions from both historic (1951–2005, "HIST") and RCP85 scenarios (2006–2100). The atmospheric fields are given as anomalies, using as a general reference state the climatology from the common period 2001–2020. For the GCM/RCM simulations, for which the *simulated* climatology is taken as reference, the corresponding sections from HIST (2001-2005) and RCP85 (2006-2020) are concatenated.

2.2 CatRaRE

We use the catalogue of radar-based heavy rainfall events (CatRaRE, Lengfeld et al., 2021), which defines heavy rainfall based on the exceedance of thresholds related to warning level 3 (roughly 5-year return level[1]) of Germany's national meteorological service (Deutscher Wetterdienst; DWD hereafter); it corresponds to more than 25 mm in one hour or 35 mm in six hours. Based on threshold exceedance of individual radar pixels, heavy rainfall objects are constructed that are contiguous in space and time, and for which an extremity index ($E_{T,A}$, Müller and Kaspar (2014)) is inferred that is a combined measure of area, duration and intensity. In this study, a day is labeled as *extreme* if the database contains an event of at most 9 hours duration on that day; it means that somewhere in Germany a corresponding severe weather was recorded, and the limited duration serves as a rough proxy that the event was convective.

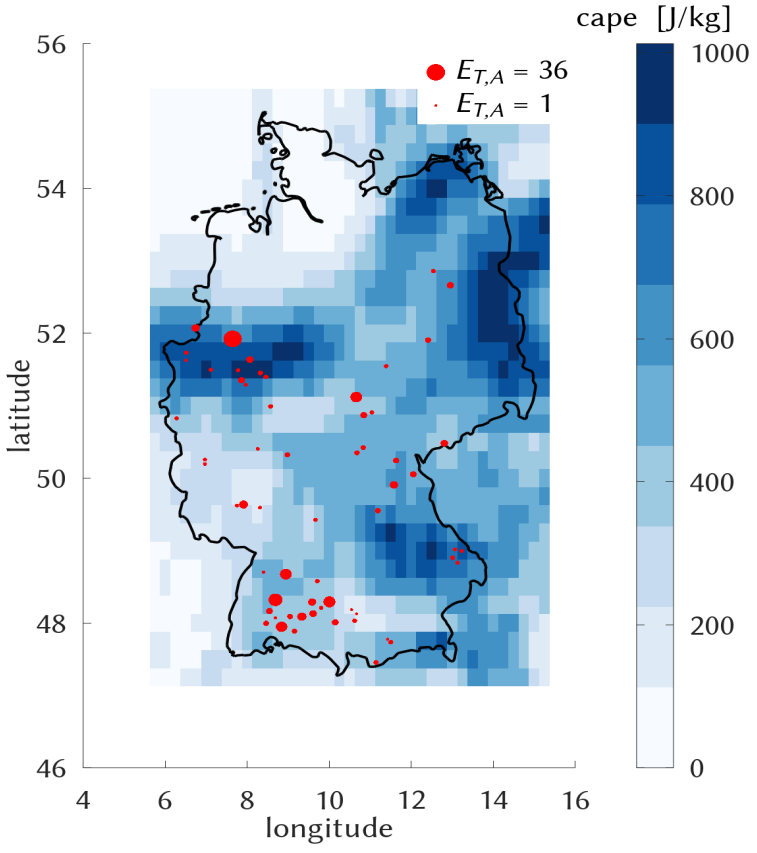

Figure 1. The conditions for cape on July 28, 2014 (blue), along with $E_{T,A}$ values of corresponding CatRaRE events of ≤9h duration (dots).

On average, 51% of the (May–Aug) days see an extreme event somewhere in Germany, which means that, although CatRaRE events are locally rare by definition, the main classification task (event vs. no event in Germany) is quite balanced. Mainly for later use we counter any potential class imbalance nevertheless, and employ a rather simplistic

[1] Given that of the total of 175200 = 20×365×24 hours from 2001 to 2020, about 27000 are listed as extreme, the likelihood of seeing any extreme event in Germany is p_G = *27000/175200 = 15%*. The average size (in pixels) of a CatRaRE event is *a=133*, while all of Germany covers a_G=900×1100 = 990000 pixels. If all CatRaRE events can be taken as independent, then the probability of an event per pixel is $p = 1-(1-p_G)^{a_G/a} = 2.25\times10^{-5}$, which roughly corresponds to a return period of 5 years.

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

oversampling approach by populating the minority class with random duplicates of that class until that class is no longer minor.

The ERA5 grid is shown in Figure 1, along with the average cape values for 28 July 2014. It was a day with particularly strong atmospheric convectivity, which led to several severe rainfall events all over Germany, as monitored by CatRaRE, so that the day is labeled as extreme. Two active regions are visible, one in the Southwest and one in the central West. There, in the city of Münster, occurred the most disastrous event, with one station recording as much as 292 l/m² within 7 hours (Spekkers et al., 2017) The surrounding cape grids show values > 600 J/kg, similar to other areas in Germany (SE, NE).

2.3 Conventional ("Shallow") and Deep Learning models

Table 1. The Shallow-Learning methods.

	abbr.	note	
Lasso regression	LASSO	cross-validated penalty	(McIlhagga, 2016)
random forests	TREE	50 trees	(Jekabsons, 2016)
shallow neural nnet	NNET	2 hidden layers with 7 and 3 neurons	Octave
logistic regression	NLS	nonlinear least squares	Octave

As competitive benchmarks to DL models, we employ four shallow statistical models: Lasso logistic regression (LASSO), random forests (TREE), and a simple neural net with 2 hidden layers (NNET); all of these are applied with and without Empirical Orthogonal Functions (EOF) orthogonalization; more details are listed in Table 1 and in the source code mentioned at the end. The architectures of the selected DL models are almost exclusively based on *convolutional neural networks* (CNNs), a concept that was introduced with the famous LeNet-5 model of (LeCun et al., 1989) for the classification of handwritten zip codes. Besides LeNet-5 we use the network architectures AlexNet, ALL-CNN, GoogLeNet, DenseNet, and ResNet. These were created for the classification of digitized images, such as the CIFAR-10 set with 32×32 image resolution and 10 classes or ImageNet with 256×256 images covering 1000 classes, and regularly used in annual image classification contests since about 2010 (Krizhevsky et al., 2017). Along with these come two quite simplistic benchmark networks, *Simple* representing a single convolutional and a dense layer, and Logreg with just one single dense layer; details are provided by Table 2 and the SI. This provides a fairly comprehensive selection from the most simple to highly sophisticated networks. The corresponding model implementations can be inspected at https://gitlab.dkrz.de/b324017/carlofff. Training and deployment of DL models is performed using the *Caffe* framework with its Octave interface (https://github.com/BVLC/caffe).

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

Table 2. The Deep-Learning architectures. The number of classes pertains to the reference study.

	Year	resolution	layers[2]	Reference	Original classes
LeNet-5	1989	28×28	4	(LeCun et al., 1989)	10
AlexNet	2012	227×227	8	(Krizhevsky et al., 2017)	1000
CIFAR-10	2014	32×32	4	(Krizhevsky et al., 2017)	10
ALL-CNN	2014	32×32	9	(Springenberg et al., 2014)	10
GoogLeNet	2014	224×224	76	(Szegedy et al., 2015)	1000
ResNet	2016	32×32	22	(He et al., 2016)	10
DenseNet	2016	32×32	159	(Huang et al., 2017)	10
Simple		32×32	3	this paper	2
Logreg		32×32	1	this paper	2

Compared to the original DL classification tasks in the literature, with e. g. 1000 classes for AlexNet and GoogLeNet, cf. Table 2, our classification in its initial form is just binary, so naturally some of the network and solver parameters had to be adjusted. A crucial "hyperparameter" is the size of the training and testing batches *(batch_size* in Caffe*)*, which had to be lowered for the broader and deeper networks. Another parameter is maximum iteration (*max_iter*); unless that number is reduced drastically the optimization would enter a runaway overfitting process whose emergence is barely visible. The learning rate decay policy *poly*, which basically required a single parameter *power*, helped to steer the learning process in a parsimonious way; it was used for all DL solvers. All adjusted parameters are listed in Table S1 from the SI.

Because DL optimization generally uses a stochastic gradient descent algorithm and is therefore not fully deterministic, we use an ensemble of 20 DL optimization runs. This ensemble, too, is informative about network convergence, and in some cases even reveals potential for refined parameter tuning. All relevant details are described in the SI, section 2.

The predictor fields of cape, tcw, and cp are taken as three 'color channels' (RGB) of an image sequence. Because the image resolution differs between the networks, varying from 28×28 pixels for LeNet-5 to 227×227 pixels for AlexNet, a regridding of the fields is required to match the resolution of the original model, cf. Table 2. Except for LeNet-5, this represents an upsampling so that the pattern itself (its shape) enters the DL essentially unchanged (and the LeNet-5 resolution is sufficiently similar). EOF truncation was consequently not applied to the DL models.

[2] We only count convolutional and fully connected (inner product) layers

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

2.4 Calibration, Validation

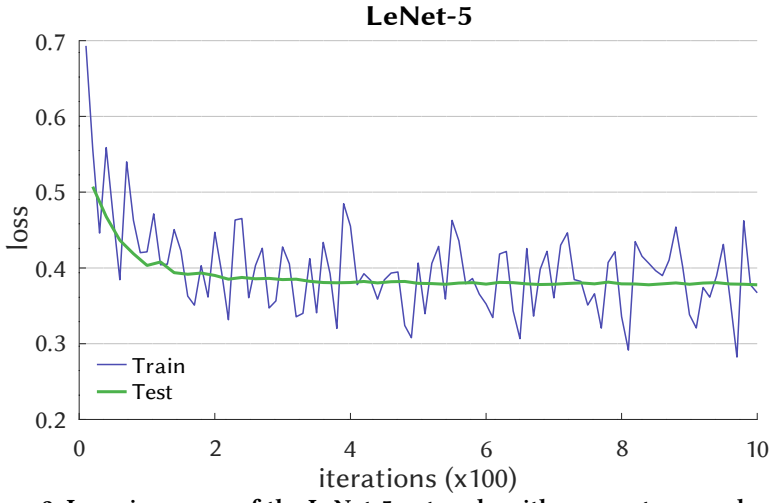

Figure 2. Learning curve of the LeNet-5 network, with crossentropy as loss. Iterations indicate the number of batch passes (batch size 100).

The full period from 2001 to 2020 amounts to a total of 2460 days, which we split into a calibration (train) and validation (test) period of 2001–2010 and 2011–2020, respectively. For the DL training, cross-entropy is used as a loss function. As evaluation measure the Equitable Threat Score (*ETS*, syn. Gilbert Skill Score) is used. ETS measures the rate of correctly forecast extremes relative to all forecasts except majority class hits, and adjusted for random hits. We note that the validation data are not completely independent of the DL models. Because they have been used for inspecting the learning curves and their convergence, there is a slight chance that the validation scores may reflect sampling properties and would therefore not generalize. On the other hand, the tuning goal was to achieve reasonable convergence of the loss function and not to minimize its value. Therefore, we are confident that overfitting is reasonably limited.

3 Results and discussion

Convergence of the DL model optimization is exemplarily shown in Figure 2, which depicts the loss function (crossentropy) during the learning and testing iterations. LeNet-5 follows a typical path of learning progress, with variable but decreasing loss for the training phase that is closely and smoothly traced by the testing phase, the latter leveling out somewhat below a loss of 0.4. The learning curves of the other networks look similar but with different absolute losses, and are shown in Figure 3. It is noticeable that e. g. ResNet converges after only 40 iterations whereas AlexNet and ALL-CNN require, respectively, 500 and 1000 iterations. Also note that the simpler networks such as Simple, Logreg, and CIFAR-10 remain stable after reaching convergence while, what is not shown in the Figure, the more complex networks AlexNet, GoogLeNet and ALL-CNN do not and start to diverge, indicative of overfitting.

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

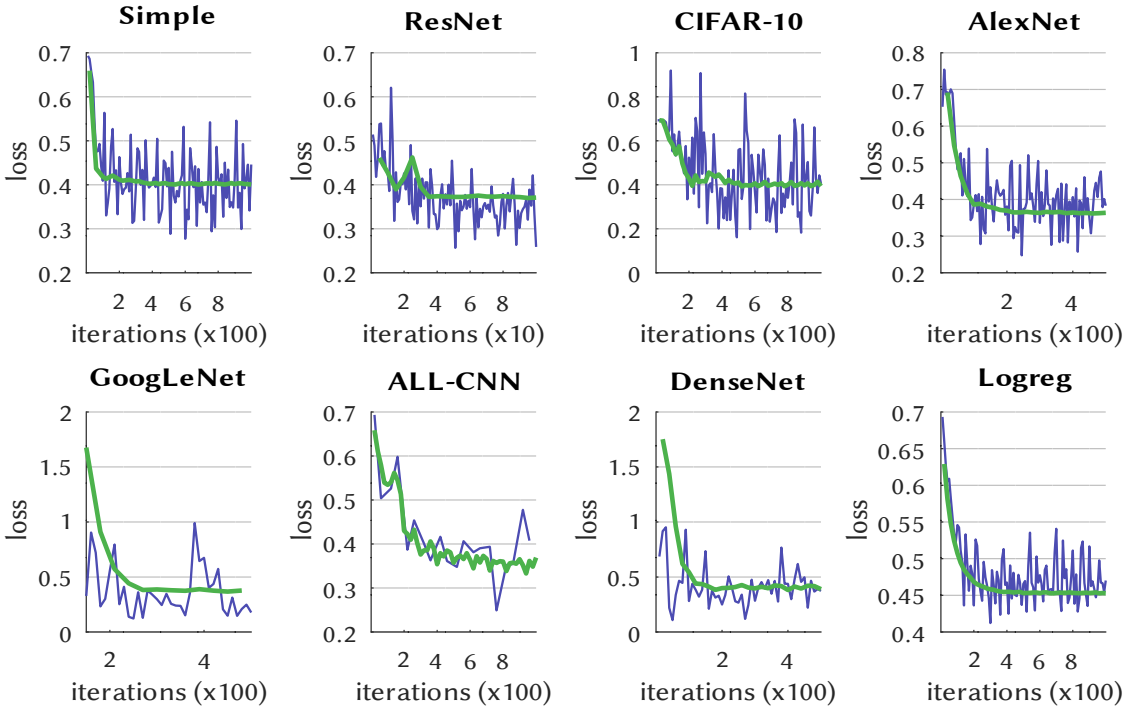

Figure 3. As Figure 2, for the other DL networks.

Overall model performance when driven by ERA5 fields from the validation period 2011–2020 is shown in Figure 4. First, it demonstrates the positive effect of using cape as a predictor, which improves skill across all models. Another distinction for the classical ("shallow") methods is the use of an EOF reduction of the predictor fields prior to the model fit; except for the (shallow) neural net the effect is positive. The best overall performance is achieved by the ALL-CNN network with a mean *ETS* of *0.52*, followed by Random Forests (TREE) with *ETS = 0.51*. For logistic regression (NLS), EOF reduction is indispensible as it otherwise leads to heavy overfitting; the neural net (NNET), on the other hand, profits from using the original instead of the reduced fields as predictors. The scatter of DL model skill, crossentropy vs. ETS, that is visible especially for the more complex models is indicative of some residual underfitting that we have not been able to resolve. But the cloud tilt is obvious, with more variation along the ETS axis. That this is not a simple scaling issue can be seen for the Logreg network, whose optimized crossentropy values, unlike the other networks, show virtually no variation compared to the ETS. Crossentropy as a loss function, so it appears, sufficiently dictates unique convergence for the training phase, but apparently does not sufficiently constrain the models to make good predictions for the testing phase. Note that all DL results are, technically, stochastic due to the stochastic nature of the optimizer. We are not aware of a more systematic discussion in the DL community addressing this kind of uncertainty, cf. e. g. Kratzert et al. (2019). It is therefore somewhat unclear how to interpret the role of, for example, the one ResNet run with *ETS ~ 0.54* that marks the best result of all. In the following DL applications the ETS-optimal ensemble member is used.

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

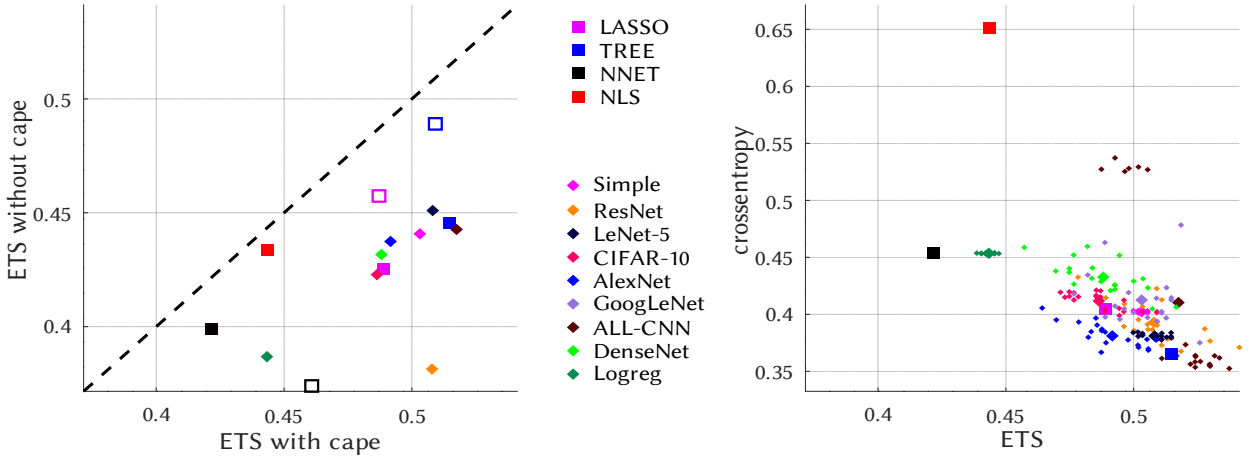

Figure 4. Model performance for the validation period 2011–2020. Left: ETS with and without cape as a predictor. Right: Relation between ETS and crossentropy (both with cape). Squares depict Shallow, diamonds Deep models. Unfilled markers in the left panel symbolize no EOF truncation. NLS without EOF truncation is outside of range.

We now apply the trained models to the observed (reanalyzed) and simulated atmospheric fields. It means we obtain for each summer day from the corresponding atmospheric model period a prediction expressing the probability of a CatRaRE-type event happening somewhere over Germany. Starting with the ERA5 reanalyses, we check whether the July 2014 event is captured by the ERA5 fields. Figure 5 shows a typical probability forecast from the DL model LeNet-5. During the days in late July of 2014, there is permanent convective activity over Germany. LeNet-5 shows near-certainty predictions for events to occur, including the July 29 extreme event. Sporadic periods of little activity are also well reflected by LeNet-5.

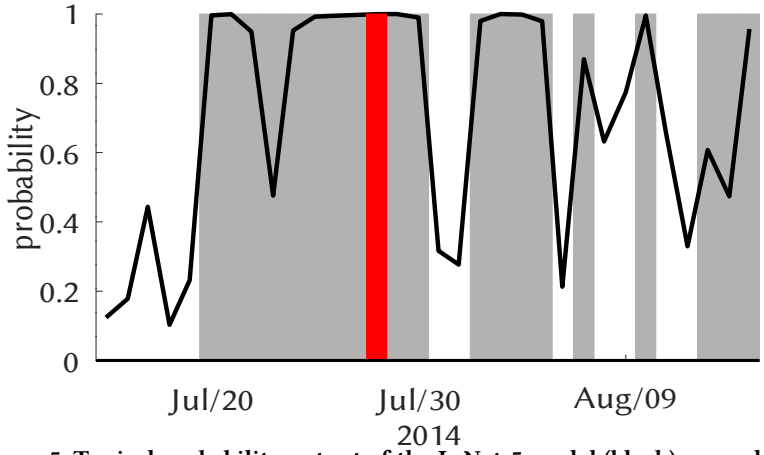

Figure 5. Typical probability output of the LeNet-5 model (black) around the July 2014 event (red); other events are gray.

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

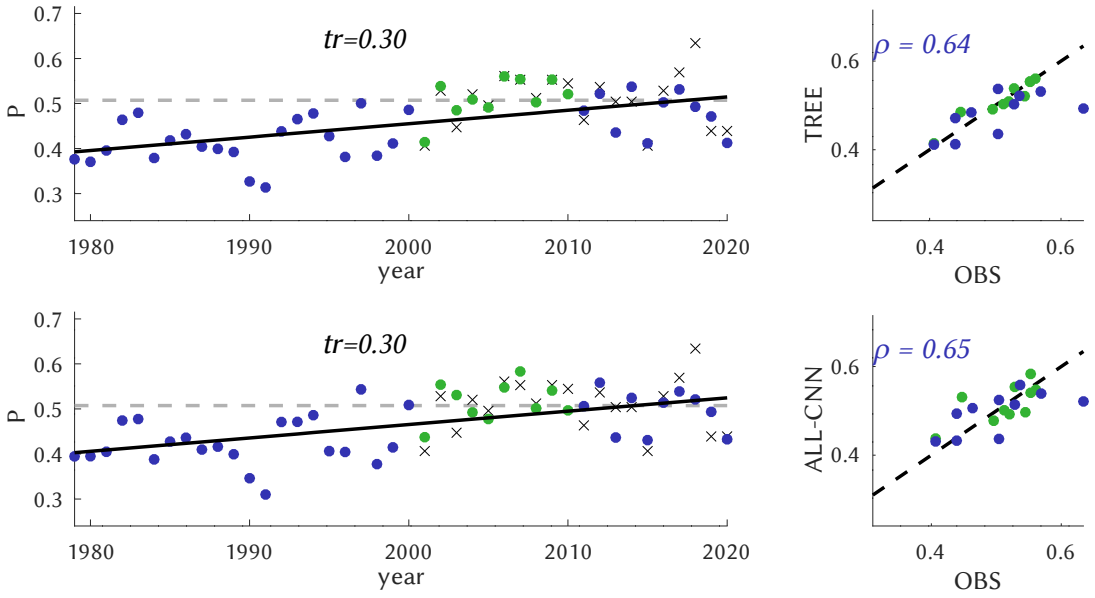

Figure 6. Annual values of the probability P of CatRaRE-type events, as observed (crosses) or simulated from ERA5 (dots), using TREE (top) and ALL-CNN (bottom); the calibration period is marked as green and the rest as blue. The full ERA5 time period reveals a significantly positive trend for both models, displayed as ΔP/100y; observed 2001–2020 climatology (gray dashed) is given for reference. The scatterplots on the right-hand side depict the same data as a scatterplot against observations, with correlations for the validation period.

For a broader temporal picture, we form annual (i. e. May–Aug) averages of the daily probabilities, and display the entire reanalysis period (1979–2020) in Figure 6. The classification is obtained by the best-scoring models ALL-CNN and TREE. The observed CatRaRE climatology (2001–2020) shows a mean daily probability of 0.51, and it is well reproduced by both models. Both models, moreover, reveal a significantly positive trend, with probabilities about 0.1 higher at the end of the period. (A linear trend is obviously only partly meaningful for a bounded quantity such as probability, but we use it here nevertheless.) Annual correlations are stronger for ALL-CNN (0.65 compared to 0.64 for TREE); corresponding plots for all other models are all very similar and are, for completeness, shown in Figs. S3 and S4; see also Table 3. Interestingly, of all models the most simple one, NLS, reveals the highest annual correlation of 0.69 with observations.

Now we analyze the CatRaRE classifications for the simulated atmospheres from past to future (1951–2100). Again, we first turn to the overall best performing models ALL-CNN and TREE, as shown in Figure 7. For TREE there is a noticeable negative bias of CatRaRE probabilities for the simulated future (2006–2100); ALL-CNN appears to be relatively unbiased. The TREE trends are not significant, while ALL-CNN exhibits significantly positive trends for both HIST *(0.12/100y)* and RCP85 *(0.06/100y)*. For the other methods the results are similar, as shown in Figs. S5 and S6, and listed in Table 3. It shows that essentially all methods consistently produce similar results, with slight variations in skill, bias, and trend. The obtained trends for the ERA5-derived CatRaRE probabilities are all fairly large and significantly positive. Interestingly, like in Figure 7, HIST trends are significantly positive for all DL models but from the Shallow methods only for TREE. Almost

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

all RCP85 trends are significantly positive, but their size is roughly half of the HIST trends. This may have to do with the limited probability domain ([0 1]) and a corresponding saturation towards the maximum.

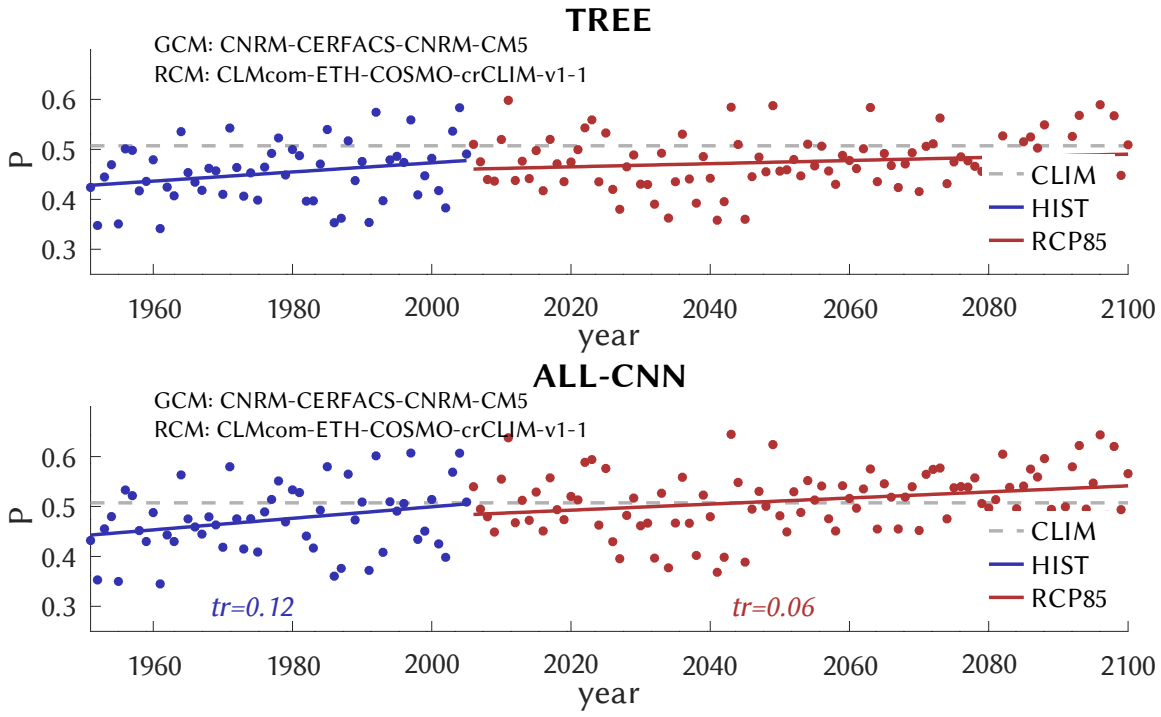

Figure 7. Similar to Figure 6, as simulated by GCM/RCM, for historic (blue) and future (red) emissions. For reference, the observed 2001–2020 climatology is also shown (CLIM, gray dashed).

4 Conclusions

205 We have classified ERA5 fields of atmospheric convectivity with respect to the occurrence of heavy rainfall events over Germany (based on the recently published CatRaRE catalog), using an array of classical ('shallow') and deep learning methods. The methods ranged from very basic logistic functions to shallow neural nets, random forests (TREE) and other machine learning techniques, including the most complex deep learning (DL) architectures that were available to us. Because of the rapid progress in DL, it still means we are at least 5 years behind the state-of-the-art. Of the classical
210 schemes, TREE performed best with an ETS score of 0.51 for the independent decade 2011–2020. Of the DL schemes, which have a stochastic component from their stochastic gradient optimizer, the overall best network was ALL-CNN with *ETS = 0.52*, but some runs of ResNet even approached ETS scores of 0.54.

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

Table 3. Summary table of ETS, trends and correlations for all methods. Significant trends are boldface. For the DL methods, the ETS ensemble mean and max is shown. Best ETS-scoring methods are blue.

model	ETS		model (ERA5) ↔ OBS annual correlation	centennial increase		
	mean	max		ERA5 (1979–2020)	HIST (1951–2005)	RCP85 (2006–2100)
LASSO	0.49		0.54	**0.22**	**0.10**	**0.07**
TREE	0.52		0.64	**0.30**	0.09	0.03
NNET	0.42		0.64	**0.24**	0.10	**0.07**
NLS	0.44		0.69	**0.28**	0.11	0.04
LeNet-5	0.51	0.52	0.58	**0.22**	**0.10**	**0.07**
AlexNet	0.49	0.52	0.60	**0.25**	**0.11**	**0.06**
CIFAR-10	0.49	0.51	0.40	**0.24**	**0.11**	**0.09**
ALL-CNN	**0.52**	0.54	0.65	**0.30**	**0.12**	**0.06**
GoogLeNet	0.50	0.53	0.49	**0.25**	**0.11**	**0.06**
ResNet	0.51	**0.54**	0.62	**0.26**	**0.11**	0.03
DenseNet	0.49	0.52	0.54	**0.32**	**0.12**	**0.07**
Simple	0.50	0.51	0.46	**0.22**	**0.10**	**0.08**
Logreg	0.44	0.45	0.54	**0.21**	**0.10**	**0.11**

The classificators were then applied to corresponding CORDEX simulations of present and future atmospheric fields. The resulting probabilities of CatRaRE-type extreme events were increasing during the ERA5 period and also for the historic and future simulations, almost independent of the methods used. Measured as centennial change, ERA5-generated probability increases by about 0.2, and this number is roughly halfed for the historic and once more halfed for the future CORDEX period. It remains unclear whether the smaller HIST rates have a real physical origin or derive from modeling inadequacies; the smaller RCP85 rates may partly be explained by a saturation effect towards maximum probability. That all probabilities increase is to be expected and in line with common wisdom of current climate research (cf. Figure SPM.6, Masson-Delmotte et al., 2021).

Compared to other classification problems such as the notorious image classification contest ImageNet, our setup of a binary classification is quite simple. One must keep in mind, however, that the very design of CNNs, with their focus on 'features' of colored shapes (objects), is modeled along the lines of ImageNet and relatives. Applying a CNN to other, not

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

object-like 'images' (blurred boundaries and colors) is not guaranteed to work out of the box. But it does, as we have seen, with only moderate adjustments. The main difficulty here was to understand just how much quicker the more complex models would learn, so that we had to shorten their learning period considerably to avoid overfitting.

Our study is meant as a starting point for a number of refinements, with the ultimate goal of classifying and projecting impact-relevant convective rainfall events for as small a region as the setting allows. So far the only criterion to isolate convective events from the CatRaRE database was their duration (here 9 hours). By considering more than two classes, e. g. by introducing more regional and temporal detail, or more levels of intensity, the full power of CNNs, and here perhaps of ALL-CNN or ResNet, could be exploited. The atmospheric predictor fields, likewise, were so far relatively simple: with local indicators of convectivity (cape, tcw, cp) whose effect can mostly be understood on a gridpoint level, the underlying statistical problem is, except for the EOF filters, essentially univariate. Using truly multivariate, pattern-based atmospheric predictors, such as moisture convergence or vorticity, can foster the performance especially of CNNs with their feature extracting capabilities. It is hoped that with all these refinements the DL methods, which are designed to handle considerably more complex classification targets, remain sufficiently reliable.

Getting back to the initial question, our conclusions entail in passing that for this study, like for so many others, machine learning methods are surpassing the conventional ('shallow') statistical toolbox. It will be interesting to see whether this also applies to state-of-the-art dynamical models. In other words, how does the development of convection-permitting dynamical models (e. g. Kendon et al., 2021) compare to DL-based convection schemes (e. g. Pan et al., 2019)? And why should their integration not offer the best of both worlds in one (Wang and Yu, 2022; Willard et al., 2022)?

5 Code availability

The relevant code underlying this paper can be found at https://gitlab.dkrz.de/b324017/carlofff.

6 Acknowledgements

We enjoyed fruitful discussions with Georgy Ayzel. The study was funded via the "ClimXtreme" (sub-project CARLOFFF, grant number 01LP1903B) by the German Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) in its strategy "Research for Sustainability" (FONA).

7 References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L.: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, Journal of Big Data, 8, 53, https://doi.org/10.1186/s40537-021-00444-8, 2021.

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

Bianco, S., Cadene, R., Celona, L., and Napoletano, P.: Benchmark Analysis of Representative Deep Neural Network Architectures, IEEE Access, 6, 64270–64277, https://doi.org/10.1109/ACCESS.2018.2877890, 2018.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., and Yacalis, G.: Could Machine Learning Break the Convection Parameterization Deadlock?, Geophysical Research Letters, 45, 5742–5751, https://doi.org/10.1029/2018GL078202, 2018.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep learning, MIT press, 2016.

Ham, Y.-G., Kim, J.-H., and Luo, J.-J.: Deep learning for multi-year ENSO forecasts, Nature, 573, 568–572, https://doi.org/10.1038/s41586-019-1559-7, 2019.

Ham, Y.-G., Kim, J.-H., Kim, E.-S., and On, K.-W.: Unified deep learning model for El Niño/Southern Oscillation forecasts by incorporating seasonality in climate data, Science Bulletin, 66, 1358–1366, https://doi.org/10.1016/j.scib.2021.03.009, 2021.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778, https://doi.org/10.1109/CVPR.2016.90, 2016.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., and Schepers, D.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, 2020.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q.: Densely Connected Convolutional Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2261–2269, https://doi.org/10.1109/CVPR.2017.243, 2017.

Jacob, D., Teichmann, C., Sobolowski, S., Katragkou, E., Anders, I., Belda, M., Benestad, R., Boberg, F., Buonomo, E., Cardoso, R. M., Casanueva, A., Christensen, O. B., Christensen, J. H., Coppola, E., De Cruz, L., Davin, E. L., Dobler, A., Domínguez, M., Fealy, R., Fernandez, J., Gaertner, M. A., García-Díez, M., Giorgi, F., Gobiet, A., Goergen, K., Gómez-Navarro, J. J., Alemán, J. J. G., Gutiérrez, C., Gutiérrez, J. M., Güttler, I., Haensler, A., Halenka, T., Jerez, S., Jiménez-Guerrero, P., Jones, R. G., Keuler, K., Kjellström, E., Knist, S., Kotlarski, S., Maraun, D., van Meijgaard, E., Mercogliano, P., Montávez, J. P., Navarra, A., Nikulin, G., de Noblet-Ducoudré, N., Panitz, H.-J., Pfeifer, S., Piazza, M., Pichelli, E., Pietikäinen, J.-P., Prein, A. F., Preuschmann, S., Rechid, D., Rockel, B., Romera, R., Sánchez, E., Sieck, K., Soares, P. M. M., Somot, S., Srnec, L., Sørland, S. L., Termonia, P., Truhetz, H., Vautard, R., Warrach-Sagi, K., and Wulfmeyer, V.: Regional climate downscaling over Europe: perspectives from the EURO-CORDEX community, Reg Environ Change, 20, 51, https://doi.org/10.1007/s10113-020-01606-9, 2020.

Jekabsons, G.: M5PrimeLab: M5'regression tree, model tree, and tree ensemble toolbox for Matlab/Octave, 2016.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, New York, NY, USA, 675–678, https://doi.org/10.1145/2647868.2654889, 2014.

Kendon, E. J., Prein, A. F., Senior, C. A., and Stirling, A.: Challenges and outlook for convection-permitting climate modelling, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379, 20190547, https://doi.org/10.1098/rsta.2019.0547, 2021.

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet classification with deep convolutional neural networks, Commun. ACM, 60, 84–90, https://doi.org/10.1145/3065386, 2017.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D.: Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation, 1, 541–551, https://doi.org/10.1162/neco.1989.1.4.541, 1989.

Lengfeld, K., Walawender, E., Winterrath, T., Weigl, E., and Becker, A.: CatRaRE_W3_Eta_v2021.01: Catalogues of heavy precipitation events exceeding DWD's warning level 3 for severe weather based on RADKLIM-RW Version 2017.002: Parameter and polygons of heavy precipitation events in Germany (2021.01), https://doi.org/10.5676/DWD/CATRARE_W3_ETA_V2021.01, 2021.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, Ö., Yu, R., and Zhou, B. (Eds.): Summary for policymakers, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 3–32, https://doi.org/10.1017/9781009157896.001, 2021.

McIlhagga, W.: penalized: A MATLAB Toolbox for Fitting Generalized Linear Models with Penalties, Journal of Statistical Software, 72, 1–21, https://doi.org/10.18637/jss.v072.i06, 2016.

Müller, M. and Kaspar, M.: Event-adjusted evaluation of weather and climate extremes, Natural Hazards and Earth System Sciences, 14, 473–483, https://doi.org/10.5194/nhess-14-473-2014, 2014.

O'Gorman, P. A. and Dwyer, J. G.: Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events, Journal of Advances in Modeling Earth Systems, 10, 2548–2563, https://doi.org/10.1029/2018MS001351, 2018.

Pan, B., Hsu, K., AghaKouchak, A., and Sorooshian, S.: Improving Precipitation Estimation Using Convolutional Neural Network, Water Resources Research, 55, 2301–2321, https://doi.org/10.1029/2018WR024090, 2019.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, PNAS, 115, 9684–9689, https://doi.org/10.1073/pnas.1810286115, 2018.

Spekkers, M., Rözer, V., Thieken, A., ten Veldhuis, M.-C., and Kreibich, H.: A comparative survey of the impacts of extreme rainfall in two international case studies, Natural Hazards and Earth System Sciences, 17, 1337–1355, https://doi.org/10.5194/nhess-17-1337-2017, 2017.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M.: Striving for Simplicity: The All Convolutional Net, arXiv e-prints, arXiv-1412, 2014.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.: Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1–9, https://doi.org/10.1109/CVPR.2015.7298594, 2015.

https://doi.org/10.5194/egusphere-2022-1159
Preprint. Discussion started: 7 November 2022

Wang, R. and Yu, R.: Physics-Guided Deep Learning for Dynamical Systems: A Survey, https://doi.org/10.48550/arXiv.2107.01272, 3 March 2022.

Weyn, J. A., Durran, D. R., Caruana, R., and Cresswell-Clay, N.: Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models, Journal of Advances in Modeling Earth Systems, 13, e2021MS002502, https://doi.org/10.1029/2021MS002502, 2021.

Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V.: Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems, https://doi.org/10.48550/arXiv.2003.04919, 13 March 2022.

Winterrath, T., Brendel, C., Hafer, M., Junghänel, T., Klameth, A., Lengfeld, K., Walawender, E., Weigl, E., and Becker, A.: Radar climatology (RADKLIM) version 2017.002; gridded precipitation data for Germany: Radar-based gauge-adjusted one-hour precipitation sum (RW) (1), https://doi.org/10.5676/DWD/RADKLIM_RW_V2017.002, 2018.