Shallow and ~~D~~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 13 statistical methods, consisting of 4 conventional ~~statistical~~ ('shallow') and 9 more recent ~~machine~~ deep machine learning (~~M~~DL) algorithms~~, and~~ ; the classifiers are then applied~~y~~ ~~them~~ to corresponding fields of simulated present and future atmospheres from the CORDEX project. ~~Due to t~~The inherent uncertainty of the DL results ~~he~~ from the stochastic nature of ~~L M~~their optimization~~there is some spread in the results~~ is addressed by employing an ensemble approach using 20 runs for each~~. T~~ network. The ~~ALL-CNN network~~ shallow Random Forest method ~~*the ; 2Equitable Threat Score (ETS) of 0.5 on average, with several learning runs exceeding an bestsperform*~~performs best with an Equitable Threat Score (ETS) around 0.52~~single best result was from ResNet with ETS = 0.54~~, followed by the DL networks ALL-CNN and ResNet with an ETS near 0.48. ~~The best performing classical scheme was a Random Forest with *ETS = 0.51.*~~ Their success can be understood as a result of conceptual simplicity and parametric parsimony, which obviously best fits the relatively simple classification task. It is found that on summer days, CatRaRE-convective atmospheres over Germany occur with a probability of about 0.5. ~~Regardless of the method, the increasing trends are predicted for~~ This probability ~~type eventsof CatRaRE~~ is projected to increase~~from ,~~ , regardless of method, both in ERA5-reanalyzed ~~from the as well as~~ and CORDEX-simulated atmospher~~fields~~es: for the historical period we find a centennial increase~~.~~ of about 0.2 and for the future period of slightly below 0.1, this smaller value likely being a saturation effect for growing probabilities.

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; Schultz et al., 2021; Reichstein et al., 2019). 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 area. The events were obtained from a recently published catalog 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)).

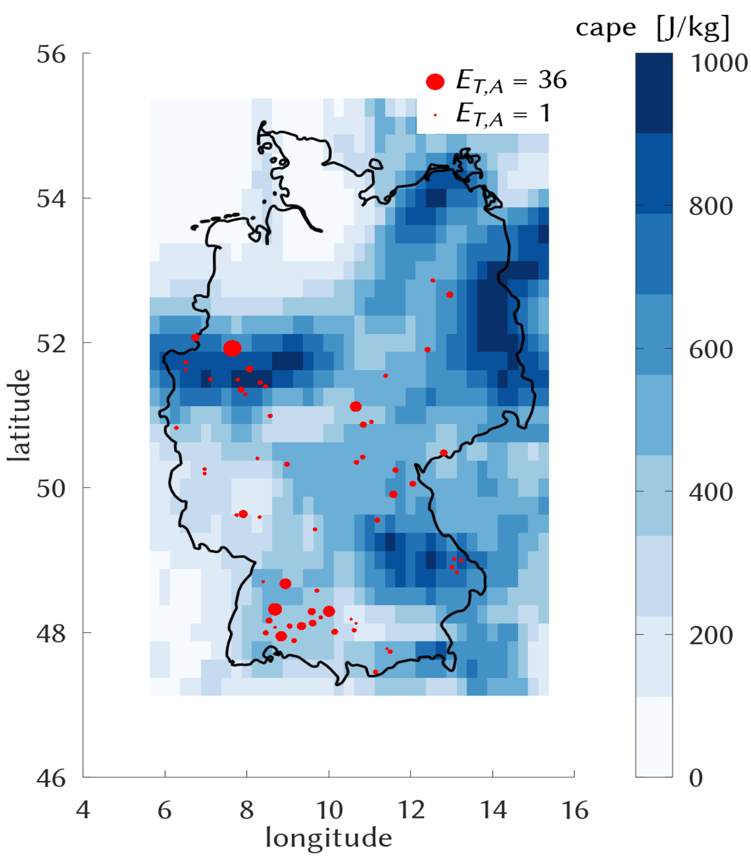

Figure 1. The conditions for cape on July 28, 2014 (blue), along with $E_{T,A}$ values of corresponding CatRaRE events of ≤9h duration (dots).

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~~ the classification performance ~~comparable to that is~~ to be at least as good as published ~~classification~~ results on those image datasets. And the prospect for using a more fine-grained analysis with more sub-regions (= more classes) should then, so we hope, be equally good. A set of methods representing state-of-the-art but conventional methods, referred to here for lack of a better expression as shallow methods, shall be used as reference.

Using standard binary skill scores, the best performing methods are applied to simulated atmospheres from the EURO-CORDEX project (Jacob et al., 2020); the predicted classification is used to estimate past and future changes in the frequency of extreme events as represented by CatRaRE. One may object that by choosing all of Germany as the (uniform) study area our approach misses important regional detail, leaving only little relevance for the local decision maker. The study is nevertheless the first of its kind to actually estimate future statistics of CatRaRE-type events ~~at all~~, and should contribute to raise awareness among researchers and decision ~~makers should read it as a form of early warning~~ makers for an impending change in these statistics. Given the wealth of methods, regional detail would at this point just add another strain to deal with, so we decided against it and do the regional assessment in an extra study afterwards.

Our focus here shall generally not be on obtaining the best result currently possible, but ~~instead of~~ rather on better understanding the influence of the 'deep' in DL with regard to performance. ~~.~~To that effect, we ~~have~~ explore~~d~~ a ~~number of conventional and newer 'shallow' methods, and compare them to a~~ selection of DL ~~networks~~ architectures that had, each in its time, ~~had~~ entered the DL arena quite spectacularly; an overview of the ~~used methods~~ architectures is given in the Supplemental Information (SI). ~~And here w~~We attempt to understand if and why they perform differently for the case of CatRaRE over Germany.

To summarize, we classify atmospheric fields of selected convectivity indices according to CatRaRE by utilizing an array statistical methods, including shallow and deep machine learning, and use those classifiers to estimate future statistics of CatRaRE-type events.

Our ~~DL~~ machine learning 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.

~~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 . 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.~~

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 classifi~~cators~~ers, 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[1]. Future atmospheric fields are obtained from the EURO-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[2]) 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 extrem~~ity~~eness 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 for that day with $E_{T,A} > 0$ and 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.

On average, 51% of the (May–Aug) days see such 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 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).

[1] In a future version, non-normality of the indices may be taken into account by using a more refined normalization (logit, probit).

[2] 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.

2.3 Conventional (“Shallow”) and Deep Learning models

Table 1. The Shallow-Learning methods.

	abbr.	note	source
Lasso regression	LASSO	cross-validated penalty (14 predictors)	(McIlhagga, 2016)
random forests	TREE	20~~5~~0 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)~~;~~. ~~a~~All of these are applied with and without Empirical Orthogonal Functions (EOF) orthogonalization, using 33, 27, and 21 EOFs for *cape*, *cp*, and *tcw*, respectively; 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).

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

	Year	resolution	layers[3]	# parameters $(\cdot 10^3)$	Reference	Original classes
LeNet-5	1989	28×28	4	400	(LeCun et al., 1989)	10
AlexNet	2012	227×227	8	60000	(Krizhevsky et al., 2017)	1000
CIFAR-10	2014	32×32	4	80	(Krizhevsky et al., 2017)	10
ALL-CNN	2014	32×32	9	1000	(Springenberg et al., 2014)	10
GoogLeNet	2014	224×224	76	10000	(Szegedy et al., 2015)	1000
ResNet	2016	32×32	22	300	(He et al., 2016)	10
DenseNet	2016	32×32	159	1000	(Huang et al., 2017)	10
Simple		32×32	3	300	this paper	2
Logreg		32×32	1	6	this paper	2

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

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. In order to stabilize the stochastic optimization, the gradient search is increasingly damped based on a factor called the base learning rate (*base_lr*)~~T~~; 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[4]. 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.

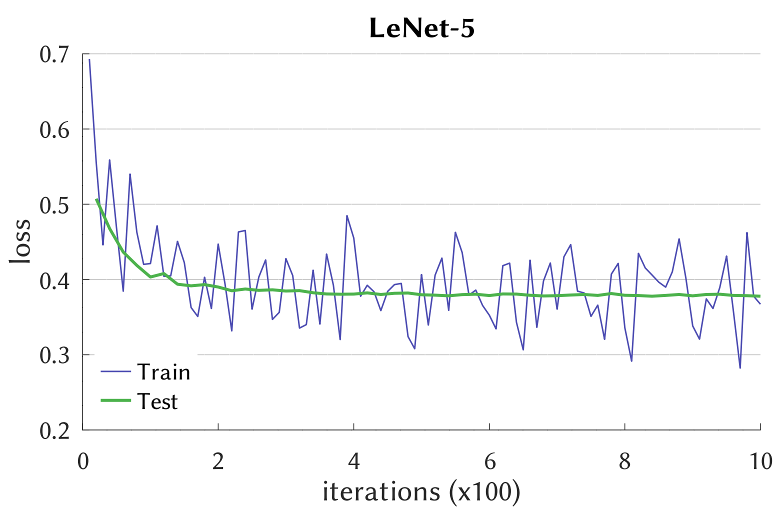

Figure 2. Learning curve of the LeNet-5 network, with crossentropy as loss. Iterations indicate the number of batch passes (batch size 100).

2.4 Calibration, Validation

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

[4] The decay at iteration *iter* is governed by the formula $base\_lr \cdot (1 - iter/max\_iter)^{power}$.

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

3.1 Network training and testing

Convergence of the DL model optimization is exempl~~arily shown~~ified in Figure 2, which depicts the crossentropy loss function ~~(crossentropy)~~ during the learning and testing (syn. calibration and validation) 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

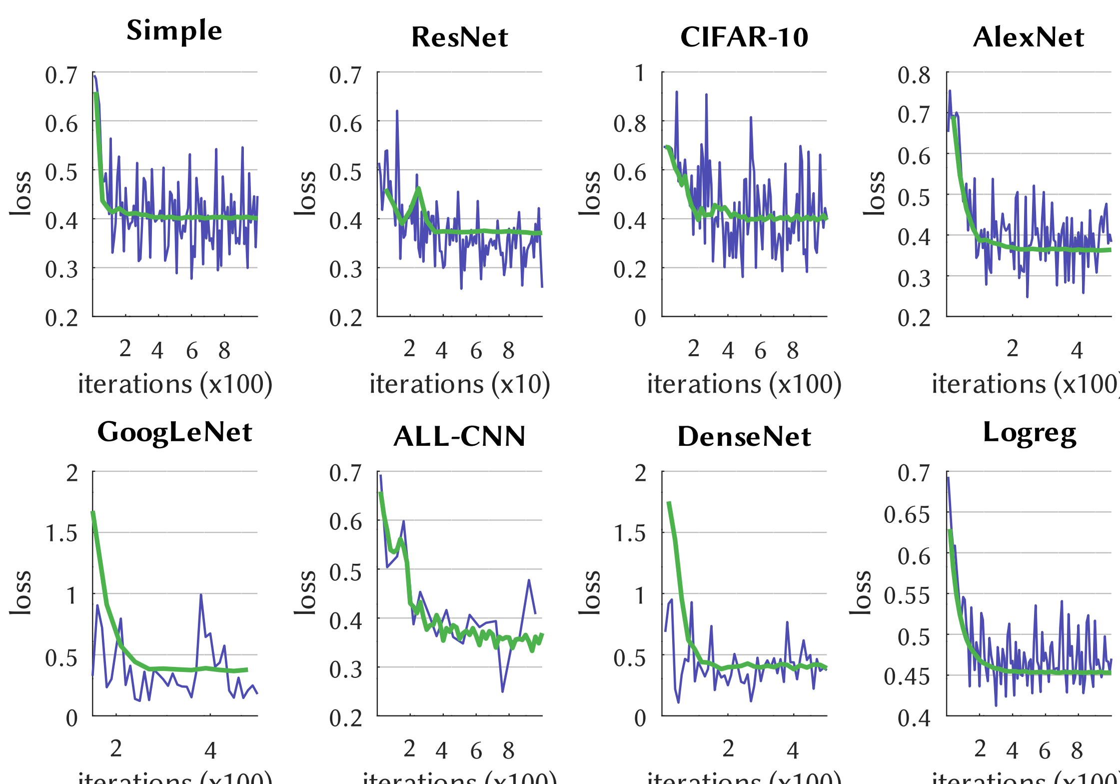

Figure 3. As Figure 2, for the other DL networks (using blue for Train and green for Test).

Figure, the more complex networks AlexNet, GoogLeNet and ALL-CNN do not and start to diverge, indicative of overfitting.

3.2 Classification performance

The probabilistic predictions are now transformed to binary (classification) predictions by choosing, from the calibration period, an optimal probability threshold for each model.

~~Overall model~~ Classification 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[5], which improves skill across all models, an exception being the poorly performing NLS model with no EOF reduction of the predictor fields ~~methods is the use of an EOF reduction of the predictor fields prior to the model fit; ")shallowclassical ("Another distinction for the .~~ ; that reduction ~~except for the (shallow) neural net the effect is positive~~obviously improves shallow model skill. The scatter of DL model skill, crossentropy ~~vs.~~versus *ETS*, is indicative of the stochastic nature that is inherent in all DL results (Brownlee, 2018; see also Kratzert et al., 2019), and uncertainty obviously grows with network complexity. The best overall

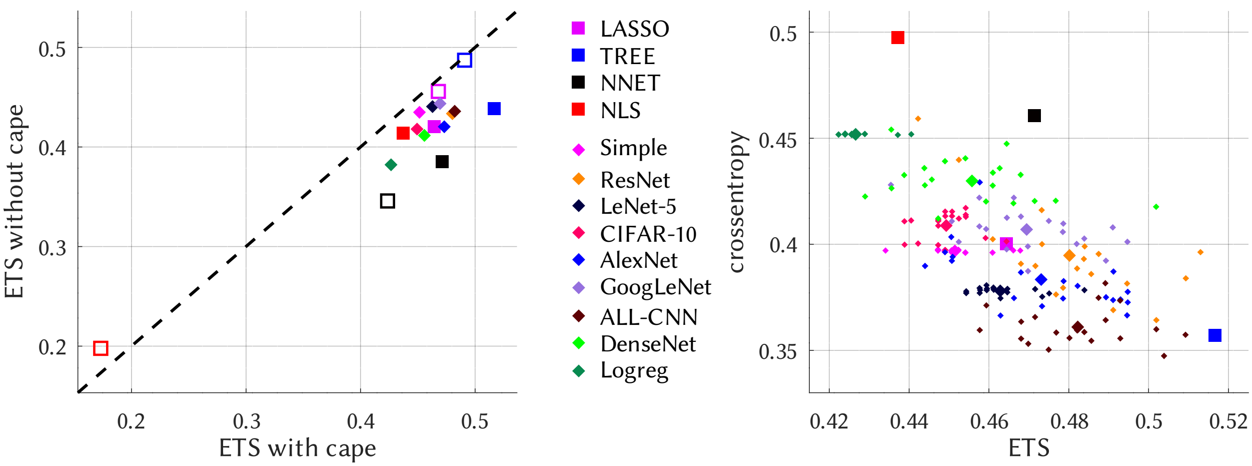

Figure 4. Model performance for the validation period 2011–2020. Left: ETS with and without cape as a predictor. Right: Relation between ETS and crossentropy (both with cape). Squares depict Shallow, diamonds Deep models. Unfilled markers in the left panel symbolize no EOF truncation. ~~NLS without EOF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

performance according to the Figure is achieved by the TREE method (~~ALL-CNN network with a mean~~ *ETS* ~~of~~ = 0.5~~2, followed by Random Forests (TREE) with ETS = 0.51~~2)~~.~~, with several of the ALL-CNN and ResNet realizations coming close, nevertheless, so that on average these turn out second. The LeNet-5, Simple and CIFAR-10 networks ~~The scatter of DL model skill, crossentropy vs. ETS, 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. indispensibleFor logistic regression (NLS), EOF reduction is~~ reveal a stretched cloud ~~tilt is obvious,~~ with ~~more~~ larger variation along the ***ETS*** axis. That this is not a simple

[5] by comparing the 3-channel predictors (cape, tcw, cp) against the two channels (tcw, cp).

scaling issue can be seen ~~for the~~ by comparing Logreg and LeNet-5~~network~~, whose optimized crossentropy values~~, unlike the other networks,~~ show virtually no variation while ~~compared to the~~ ***ETS*** varies stronger. Crossentropy as a loss function, so it appears, sufficiently dictates unique convergence for the training phase, but apparently does not ~~sufficiently~~ constrain the models enough to make good predictions for the testing phase. For logistic regression (NLS), EOF reduction is indispensable as it otherwise leads to heavy overfitting. Stochasticity is not limited to DL, it is also contained in NNET as a 'normal' neural net and, as the name suggests, random forests (TREE). ~~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. 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.~~ Like for the DL networks we form ensembles also for NNET and TREE, as further explained in the SI. And as Fig. S3 demonstrates, a second realization of the shallow and deep ensembles essentially yields similar results. In the following DL applications the ETS-optimal ensemble members ~~is~~ are used.

Differences in DL model performance ~~is~~are difficult to interpret, but a few hints may be obtained by inspecting the network architecture. Quite roughly, the width of a convolutional network represents the number of learnable features whereas the depth measures the grade of abstraction that can be formed from these features. A convective atmospheric field is, compared to a landscape with cats or dogs in it, quite simple. If a network architecture scales well this simplicity should not matter. However, very rich architectures also require a wealth of data to learn their many parameters from (14M images in ImageNet), which we do not have here. Particularly the very wide and/or deep networks such as AlexNet, GoogLeNet, or DenseNet may suffer either from inferior scaling behavior or too little data. ALL-CNN and ResNet, on the other hand, are designed particularly for simplicity and parsimony (Springenberg et al., 2014; He et al., 2016), with good performance across a broad spectrum of applications and apparently best adapted to our case.

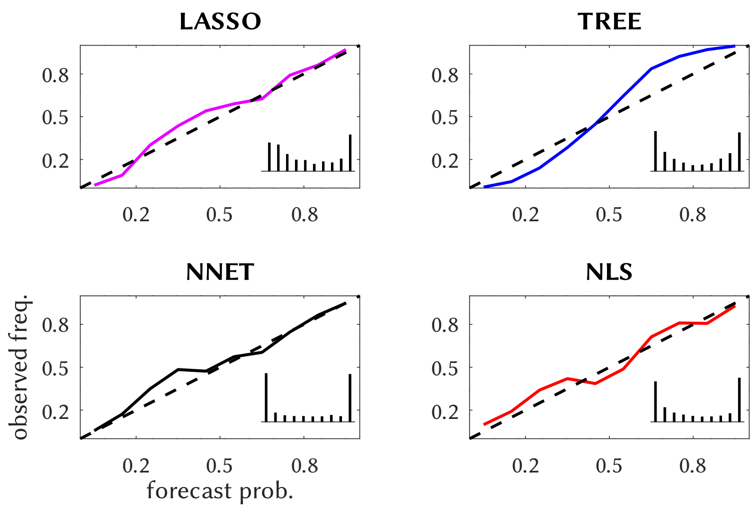

Figure 5. Reliability diagram for the shallow methods, with forecast histogram inset based on 10 bins and constant y-axis scale.

3.3 Probabilistic reliability and sharpness

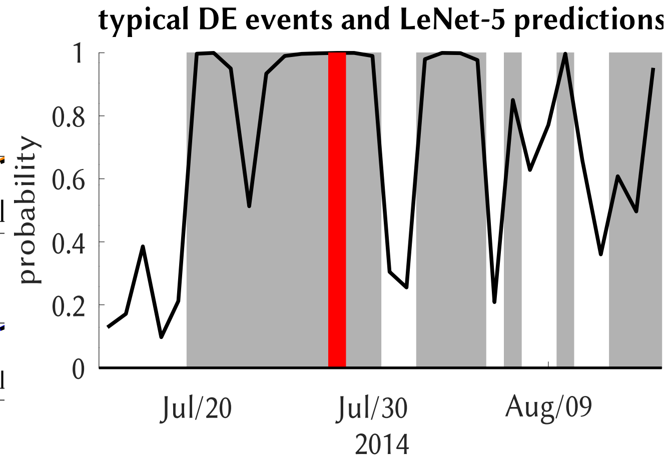

Figure 7. Typical probability output of the LeNet-5 model (black) around the July 2014 event (red); other events are gray.

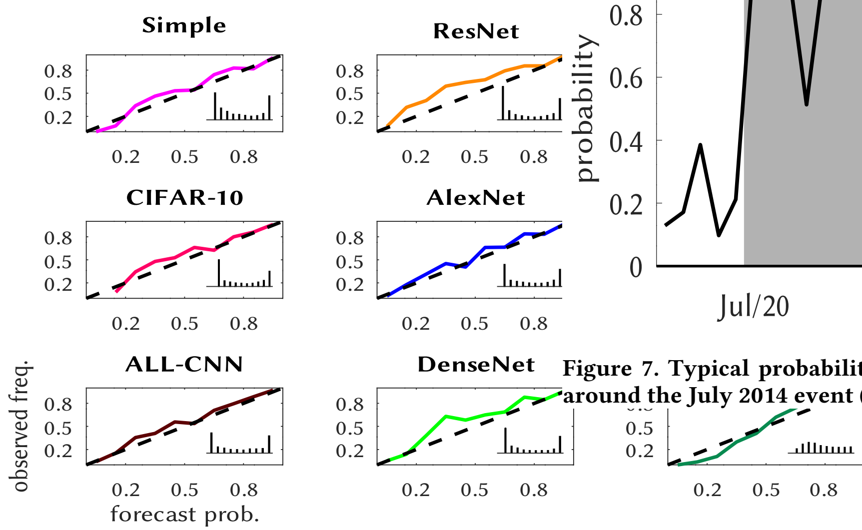

Figure 6. Like Fig. 5, for the deep methods.

We further illustrate the reliability of the probabilistic predictions by means of reliability diagrams, first in Fig. 5 for the shallow methods; ~~inset~~ these come with an inset forecast histogram,

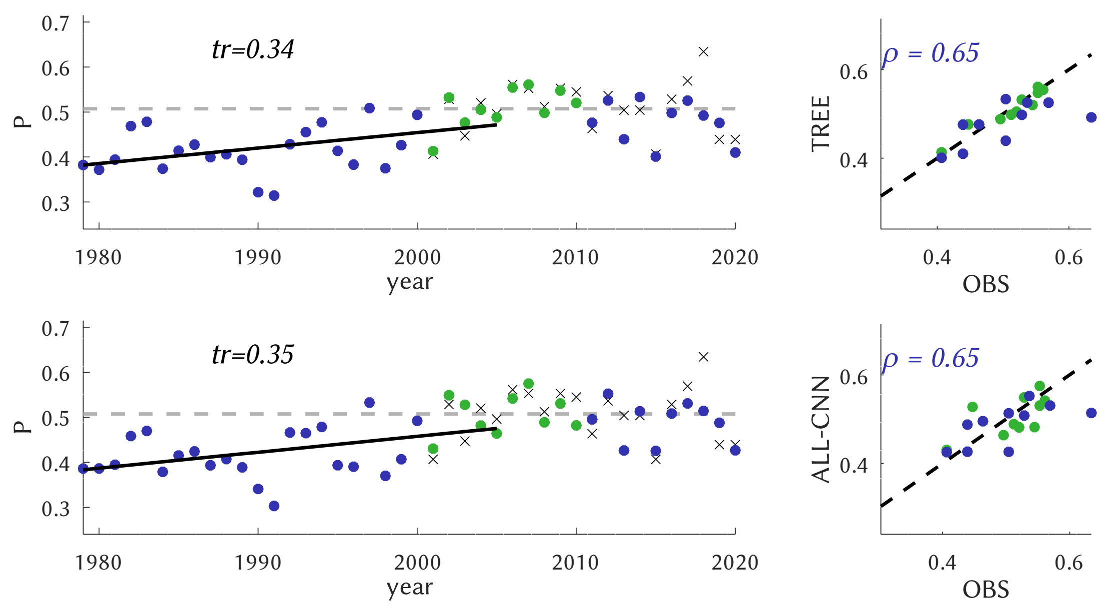

Figure 8. Annual values of the probability P of CatRaRE-type events, as observed (crosses) or simulated from ERA5 (dots), using TREE (top) and ALL-CNN (bottom); the calibration period is marked as green and the rest as blue. The 1979–2005 time period reveals a significantly positive trend for both models, 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.

displaying the relative frequencies of the delivered probabilities as a measure of sharpness of the prediction. ~~-~~The methods are quite reliable, except that TREE's lower-probability predictions occur too rarely and ~~as predicted~~ the higher ones too often. ~~A~~LASSO and TREE predictions are, as the inset shows, moderately sharp, unlike NLS and especially NNET which is almost perfectly sharp. Most of the deep methods are reliable, cf. Fig. 6, exceptions being ResNet, DenseNet and GoogLeNet, whose predictions of medium probabilities occur too often. They are also more reliable than the shallow

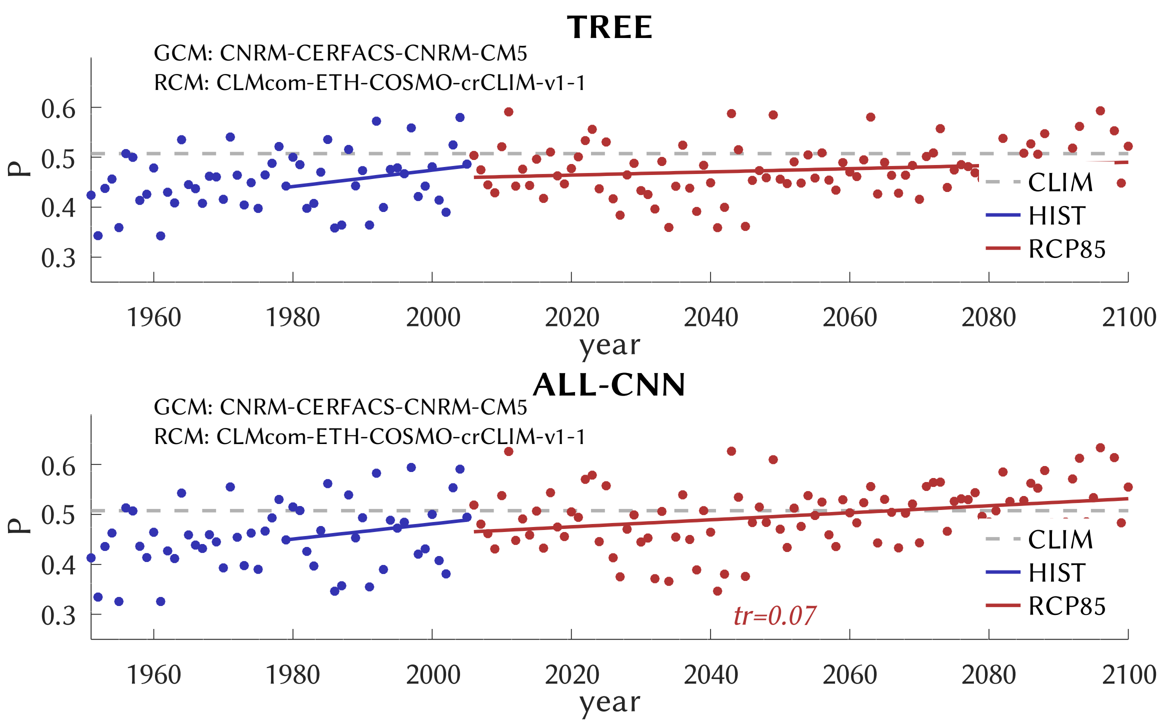

Figure 9. Similar to Figure 8, as simulated by GCM/RCM, for historic (blue) and future (red) emissions. For reference, the observed 2001–2020 climatology is also shown (CLIM, gray dashed).

methods and generally sharper, especially CIFAR-10, GoogLeNet, and DenseNet with a high load of near yes/no predictions.

3.4 Model application

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 7 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.

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 8. The classification is obtained ~~by~~from the best-scoring model~~s~~ TREE along with 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. For the period 1979–2005, which is the common period for historic reanalyses and simulations, ~~, moreover,oth modelsB~~they reveal ~~a~~ strong significantly[6] positive centennial trends of 0.34 and 0.35, respectively~~the periodhigher at the end of 0.1 about with probabilities~~. (A linear trend is obviously only partly meaningful for a bounded quantity such as probability, but we use it here nevertheless.) Annual correlations are equally strong~~er for ALL-CNN~~ (0.65 ~~compared to 0.64 for TREE~~for both); corresponding plots for all other models are ~~all very~~ altogether similar and are, for completeness, shown in Figs. S3 and S4; note, however, that the centennial trends are slightly weaker, ~~see~~ as also shown in Table 3. Interestingly, ~~of all~~ the model~~s the most simple one~~ with almost the poorest daily performance (*ETS = 0.44*), NLS, reveals the highest annual correlation of 0.6~~9~~2 with observations.

Now we analyze the CatRaRE classifications for the simulated atmospheres from past to future (1951–2100), based on HIST and RCP85. Again, we first turn to the overall best performing model~~s~~ TREE along with ALL-CNN ~~and TREE~~, as shown in Figure 9. ~~For TREE there is a noticeable negative bias of CatRaRE probabilities for the simulated future (2006–2100); ALL-CNN~~ Both appear~~s~~ to be relatively unbiased (with respect to the normal period 2001–2020) and, as Table 3 shows, the HIST simulations exhibit positive trends (1979–2005) of around 0.16 and 0.15, respectively, which amounts to only half of what was seen for ERA5~~TREE trends are not significant, while The .~~; for RCP85, only ALL-CNN exhibits a significantly positive centennial trend~~s RCP85 (both HIST (0.12/100y) and for~~ *of 0.067*~~/100y)~~. The much larger ERA5 trends as compared to HIST are actually seen only for these two methods, as for all other methods the trends for ERA5 and HIST are more similar~~are closer~~, cf. Table 3, Figs. S5 and S6. While this case may be related to the stronger annual correlations, the discrepancy between ERA5 and HIST trends for the other methods remains unclear. ~~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 9, HIST trends are significantly positive for all DL models but from the Shallow methods only for TREE. ,For the other methods the results are similar, as shown in Figs. S5 and S6 Almost all~~ The RCP85 trends are, except for TREE, significant but smaller~~, but their size is roughly half of the HIST trends. This significantly positivemay have to do with~~ , which we explain as ~~the limited~~ a saturation for growing probabilit~~y domain ([0 1]) and a corresponding saturation towards the maximum~~ies.

A final note of caution may be in place. In our modeling approach we have tacitly assumed that the learned statistical relationships remain valid when applied to previously unknown atmospheres, and remain so even when those are from a dynamical simulation or a different climate. That this may indeed cause problems became apparent when going from

[6] We us~~inge~~ a significance level of ***α=0.05*** throughout the study.

ERA5 to HIST, with average trends dropping, by reasons unknown, to almost a half. This is an epistemic problem as old as statistical climate research itself, and relates back to the concept of *perfect prognosis* (Klein et al., 1959) or in newer form to the *concept drift* in machine learning (Widmer and Kubat, 1996). There is no generally valid argument in support of the approach, and one must resort to heuristic reasoning[7]. With respect to applying simulated predictor fields (classifiers) it is usually assumed that their simulation is sufficiently reliable. And as recent analyses have shown (Kendon et al., 2021), one should indeed not be too confident in our convective classifiers *(cape, tcw, cp)* as compared to, e. g., pressure or temperature fields. With respect to a different climate, the argument is that the difference can still be seen as an anomaly from a base state and not as a shift to a wholly new climate regime, and at least for now there is little evidence for that latter case. – Given this uncertainty, the trend projections of this study, which were derived from a single climate model, are remarkably stable, indicating that progress in this direction mainly lies in the dynamical modeling of convection.

4 Conclusions

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 conventional ('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. The conventional random forest scheme ~~Of the classical schemes,~~ TREE performed best with ~~an~~ an ETS classification score ~~score of ETS~~ near 0.5~~1~~2 for the independent ~~decade~~ validation period 2011–2020, followed by the DL networks ALL-CNN and ResNet. Those schemes seem to be best adapted for the CatRaRE classification problem presented in this study~~.~~: ~~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~~TREE uses a clever bootstrap aggregating (*bagging*) algorithm over simple decision trees (200 in our case) whose generalization capacity is obviously crucial; and ALL-CNN and ResNet are networks of fairly moderate width and depth, for which training and testing performance are in balance.~~.~~

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 (mean)	model (ERA5) ↔ OBS annual correlation	centennial increase		
			ERA5	HIST	RCP85 ~~(~~2006–2100~~)~~
			~~(~~1979–2005~~)~~		
LASSO	**0.46~~9~~**	**0.54**	**0.2~~2~~5**	**0.21~~10~~**	**0.07**
TREE	**0.52~~2~~**	**0.65~~1~~**	**0.34~~1429~~**	**0.16~~09~~**	**0.03**

[7] Observational records of the relevant variables that could be used for verification are not long enough.

NNET	0.47~~24~~	0.57~~2~~	0.~~2~~32~~24~~	0.~~1~~20~~1~~	0.0~~3~~7
NLS	0.44~~4~~	0.62	0.~~27~~31	0.~~11~~21	0.05
LeNet-5	0.46~~51~~	0.58	0.27~~2~~	0.~~11~~22	0.07
AlexNet	0.47~~9~~	0.59~~60~~	0.33~~25~~	0.1~~1~~8	0.05
CIFAR-10	0.45~~9~~	0.39~~40~~	0.24~~4~~	~~0.11~~0.20	0.07
ALL-CNN	0.48~~52~~	0.65~~8~~	0.35~~26~~	0.15~~2~~	0.0~~4~~7
GoogLeNet	0.47~~50~~	0.54~~09~~	0.30~~52~~	0.~~09~~20	0.0~~4~~5
ResNet	0.48~~51~~	0.58~~62~~	0.30~~26~~	0.~~11~~20	0.04
DenseNet	0.46~~9~~	0.51~~4~~	0.33~~32~~	0.~~12~~20	0.05
Simple	0.45~~50~~	0.46	0.24~~2~~	0.~~10~~22	0.08
Logreg	0.43~~4~~	0.54	0.18~~21~~	0.~~10~~21	0.11

The classifi~~cators~~ers were then applied to corresponding CORDEX simulations of present and future atmospheric fields. The resulting probabilities of convective atmospheric fields and related CatRaRE-type extreme events were increasing during the ERA5 period and also for the historic and future CORDEX simulations, ~~almost~~ independent of ~~the~~ method ~~useds~~. This is to be expected and in line with common wisdom of current climate research (cf. Figure SPM.6, Masson-Delmotte et al., 2021). Specifically, using TREE for the historic period (1979–2005) ~~Measured as centennial change, ERA5-generated~~ the resulting probabilit~~y~~ies, measured as centennial trend, increase~~s~~ by ~~about~~ 0.2~~3~~4 for ERA5 and by 0.16 for HIST. ~~,~~ We were unable to resolve this discrepancy (which is less severe for the other methods) and its potential modeling inadequacy remains unclear. ~~and this number is roughly halfed for the historic and once more halfed f~~For the future CORDEX ~~period~~simulations we obtained a smaller but significant increase of around 0.07 for most methods~~. tIt remains unclear whether the smaller HIST rates have a real physical origin or derive from modeling inadequacies; he smaller RCP85 rates may~~ , a number that can partly be explained by a saturation effect for growing probabilities~~towards maximum probability~~. The overall tendency towards more extreme convective sub-daily events is consistent with recent estimates from Clausius-Clapeyron temperature scaling (Fowler et al., 2021) as well as from a convection-permitting dynamical climate model for Germany (Purr 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 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. That way, the usefulness of the results for decision makers in risk management could be increased substantially. 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 especially 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 at least for this study~~, like for so many others~~, ~~machine~~ deep learning methods are not surpassing the conventional ('shallow') statistical toolbox. It will be interesting to ~~see whether this also applies to~~ follow the evolution in state-of-the-art dynamical models. ~~In other words~~Specifically, 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 Author contribution

GB and MH designed the experiments and GB carried them out, developed the model code, performed the simulations and prepared the manuscript with contributions from MH.

7 Competing interests

The authors declare that they have no conflict of interest.

8[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).

9[7] Declaration

The authors declare that they have no conflict of interest.

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.

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.

Brownlee, J.: Statistical Methods for Machine Learning: Discover how to Transform Data into Knowledge with Python, Machine Learning Mastery, 291 pp., 2018.

Fowler, H. J., Lenderink, G., Prein, A. F., Westra, S., Allan, R. P., Ban, N., Barbero, R., Berg, P., Blenkinsop, S., Do, H. X., Guerreiro, S., Haerter, J. O., Kendon, E. J., Lewis, E., Schaer, C., Sharma, A., Villarini, G., Wasko, C., and Zhang, X.: Anthropogenic intensification of short-duration rainfall extremes, Nat Rev Earth Environ, 2, 107–122, https://doi.org/10.1038/s43017-020-00128-6, 2021.

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.

Klein, W. H., Lewis, B. M., and Enger, I.: Objective Prediction of Five-Day Mean Temperatures during Winter., Journal of Atmospheric Sciences, 16, 672–682, 1959.

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.

Purr, C., Brisson, E., and Ahrens, B.: Convective rain cell characteristics and scaling in climate projections for Germany, International Journal of Climatology, 41, 3174–3185, https://doi.org/10.1002/joc.7012, 2021.

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.

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, 20200097, https://doi.org/10.1098/rsta.2020.0097, 2021.

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.

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.

Widmer, G. and Kubat, M.: Learning in the presence of concept drift and hidden contexts, Machine learning, 23, 69–101, 1996.

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.