Meteorol. Z. (Contrib. Atm. Sci.), Vol. **32**, No. 6, 487–497 (published online August 17, 2023)
© 2023 The authors

Atmospheric Dynamics

A machine learning approach on the investigation of the scale dependent relation of CAPE and precipitation

Annette Rudolph[1,2*] and Peter Névir[2]

[1]Technische Universität Berlin, Berlin, Germany
[2]Freie Universität Berlin, Berlin, Germany

(Manuscript received January 27, 2022; in revised form March 14, 2023; accepted July 17, 2023)

Abstract

The temporal and spatial scale dependent relation of Convective Available Potential Energy (CAPE) and precipitation is investigated. Using the COSMO-REA6 data set, we ask which of the standard machine learning algorithms: perceptron, support vector machine, decision tree, random forest, k-nearest neighbor and a simple kept deep neural network algorithm can best relate these two variables. Then, we concentrate on decision trees and evaluate the relation of CAPE and precipitation across different scales. We investigate temporal resolutions of 1 hour to 24 hours and horizontal resolutions of 6 km up to 768 km. Regarding ten CAPE and two precipitation classes we find accuracy scores mostly of about 0.7 across all scales. Taking the Dynamic State Index (DSI) as additional predictor into account leads to an overall increase of the scores. We further introduce a theoretical relation of CAPE and precipitation based on the works of Hans Ertel (1933), which will be analyzed in future studies. Today it is natural to tackle complex atmospheric processes using machine learning methods. These data based methods are suggested as additional tool to complement the results gained by the governing equations of atmospheric motion.

Keywords: Precipitation, CAPE, DSI, decision tree

1 Introduction

Precipitation, its impact and forecast is a present topic in our daily life. But cloud physics is not fully understood leading to uncertainties in the forecast of rainfall. Especially convective precipitation can be very local and the intensity can vary even between different urban districts of one city. Regarding the larger scales, synoptic fronts can be stretched in the order 1000 kilometers. They can be detected on satellite images and their forecasts are quite good. Even though smaller convection can be detected on satellite images too, the exact location of rainfall is hard to predict. From dynamical perspective, precipitation is related to atmospheric instability that is characterized by a large vertical temperature gradient. The relation of extreme precipitation and temperature anomalies is for example shown in Müller et al. (2020). However, a more accurate parameter that takes the vertical temperature gradient into account and measures hydrostatic instability is the Convective Available Potential Energy, short CAPE, see e.g. Weisman and Klemp (1982), Holton (2004), Khouider (2019). Assuming adiabatic conditions and that there is no mixing of an air parcel with its environment during ascent, CAPE measures, how much an air parcel can be lifted and how much kinetic energy could be obtained. Let now T_v be the virtual temperature that is approximately given by

$$T_v \approx T(1 + 0.61\frac{\rho_v}{\rho}) \tag{1.1}$$

with the density of water vapor ρ_v and the density of dry air ρ, see e.g. the book of Markowski and Richardson (2011). Considering a moist air parcel, its virtual temperature is the temperature at which the total pressure and density of the theoretical dry air parcel is equal to the moist air parcel. Following Markowski and Richardson (2011) and expressing the buoyancy B as the virtual temperature perturbation of a lifted air parcel T_v' divided by the virtual temperature of the environment $\overline{T_v}$, CAPE can be defined as follows:

$$CAPE = g\int_{z_{\mathrm{LFC}}}^{z_{\mathrm{ET}}} B\,dz = g\int_{z_{\mathrm{LFC}}}^{z_{\mathrm{ET}}} \frac{T_v'}{\overline{T_v}}\,dz\,, \tag{1.2}$$

where z_{LFC} is the so-called *Level of Free Convection*, short LFC. At this height z_{LFC}, the rising air parcel becomes significantly warmer than its environment, z_{ET} is the height, where the rising air parcel has equal temperature (ET) to its environment and g is the acceleration due to gravity. The virtual temperature of the lifted air parcel is $T_v = \overline{T}_v + T_v'$.

CAPE and its relation to convection has been investigated in several studies, see for example Rennó and Ingersoll (1996), who present a theory for convection related to CAPE based on the heat engine framework, or see the work of Ramezani Ziarani et al. (2019), who consider CAPE and dew-point temperature to charac-

[*]Corresponding author: Annette Rudolph, Technsiche Universität, Berlin, e-mail: annette.rudolph@tu-berlin.de

DOI 10.1127/metz/2023/1130
Gebrüder Borntraeger Science Publishers, Stuttgart, www.borntraeger-cramer.com

terize rainfall-extreme events in the South-Central Andes. Recently, Rybka et al. (2020) investigate CAPE numerically and Polzin et al. (2022) study the relation of CAPE to the vertical velocity considering a direct Bayesian model reduction algorithm.

Even though many works have confirmed that CAPE is a useful variable to indicate convective events, also considered in climate studies, from the definition of CAPE it follows, that the values of CAPE should not be taken as exact determined numbers (Riemann-Campe et al., 2009; Adams and Souza, 2009; National Weather Service, 2023; Williams and Renno, 1993). Therefore, to study relations of CAPE and precipitation, classification algorithms seem to be a reasonable choice. Moreover, the steadily grown amount of data of increasing resolutions together with today's numerical possibilities to apply machine learning algorithms motivate to consider these tools to address the following research questions:

1. Can we use todays increasing possibilities of the applicability of machine learning algorithms to investigate the relation of CAPE and precipitation across various temporal and spatial scales?
2. Which algorithm provide the best relations of CAPE and precipitation?
3. Which benefits provide the Dynamic State Index (DSI) and the Thunderstorm Occurrence Parameter (TOP) as additional input variable?
4. Is there a theory of the relation of CAPE to precipitation?

In order to answer these questions, the paper is structured as follows. In Section 2 we summarize the machine learning algorithms and the accuracy scores that we use for our study. The accuracy score is defined as the number of correctly classified data instances over the total data. We apply these methods to data that we outline in Section 3, where we also describe the steps of preprocessing. The results of the time and spatial dependent relation of CAPE and precipitation are represented in Section 4. Furthermore, we show that taking the DSI as additional input variable into account leads to higher scores. As an alternative approach, we shortly summarize the theoretical relation of CAPE and precipitation based on the work of Hans Ertel (Ertel, 1933) in Section 4.4 and finally summarize our results in Section 5.

2 Methods

To investigate the temporal and spatial scale-dependent relation of CAPE and precipitation, we consider the following methods: classical logistic regression, perceptron algorithms, support vector machine (svm), decision tree/random forest, k-nearest neighbor and a simple deep neural network. First we give a short summary of these methods. We will use the following notation: $\boldsymbol{x} = (x_1, \ldots, x_m)$ are m CAPE categories that we relate to two precipitation classes $y = 1$ or $y = 0$, where $y = 1$ is the class of all precipitation intensity events with intensity greater than the 75th percentile and $y = 0$ represents the class of lower precipitation intensity. For all models, the vector $\boldsymbol{w} = (w_1, \ldots, w_n)$ denotes the weights for n inputs that are treated differently by the algorithms summarized in **a)–f)**. See e.g. Raschka and Mirjalili (2017) for a more detailed explanation of these methods and for the implementation with the programming language python.

a) Logistic regression is a widely used, linear classification model. Let p be the probability for $y = 1$ (heavy precipitation). The so-called logit function is defined as the logarithm of the odd ratio $p/p - 1$:

$$\mathrm{logit}(p) = \log \frac{p}{p-1} \tag{2.1}$$

Now, let $p(y = 1|\boldsymbol{x})$ the conditioned probability that a sample belongs to $y = 1$ given by its feature $\boldsymbol{x}$. The linear relation between the feature values and the log-odds is given by:

$$\mathrm{logit}(p(y = 1|\boldsymbol{x})) = \sum_{i=0}^{m} w_i x_i = \boldsymbol{w}^T \cdot \boldsymbol{x} := z \tag{2.2}$$

The inverse gives us the probability that a particular sample belongs to one of the classes. This is called the logistic sigmoid function ϕ:

$$\phi(z) = \frac{1}{e^{-z}}, \tag{2.3}$$

which has the characteristic S-Shape.

b) Perceptron algorithm is a linear classifier that is used for supervised learning of binary classifiers. Roughly speaking, it finds the best line separating two data sets. Already in 1958 the algorithm was introduced by Frank Rosenblatt. The algorithm learns a threshold function that maps the real-valued input vector (e.g. of CAPE values) $\boldsymbol{x} = (x_1, \ldots, x_n)$ to the output $f(\boldsymbol{x})$, which is either 1 (*precipitation over a a certain threshold*) or 0 (*no precipiation*). Mathematically, the function f is given by:

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} w_i \cdot x_i + b > 0 \\ 0 & \text{otherwise}\,, \end{cases} \tag{2.4}$$

where $w_i, \ldots w_n$ are the weights, n the number of inputs and b the bias.

c) Support vector machine (svm). This machine learning algorithm is designed to find a hyperplane in a N-dimensional space, where N is the number of features. The hyperplane classifies the data points. Our goal is to distinguish between two classes of precipitation intensity, where the precipitation intensity is a scalar, therefore, the hyperplane is a line. The key difference to the perceptron algorithm is that the perceptron algorithm stops after it classifies data correctly. In contrast, svm finds the best plane with the the maximum margin. This is the general objective of svm: Finding a hyperplane (i.e. a line) that has the maximum distance, or maximum margin, between data points of two classes. Obtaining a maximal margin distance leads to a more precise classification of future data points.

d) k-nearest neighbor (knn) This algorithm determines k nearest neighbors. First, using a training data set, clusters of similar characterizations are found. Each cluster is a class. A new data point is then assigned to the majority class of nearest neighbors.

e) Decision tree/random forest A decision tree works as the name indicates: Starting in a tree root, the data is recursively split. The simple decision rules of the splits are inferred from the training data. A random forest can be seen as a number of decision trees.

f) Simple Deep learning A neural network with multiple layers between the input and output layers is called a Deep Neural Network (DNN). One advantage of the use of a DNN with more than one layer is that such a network produces a nonlinear decision boundary with nonlinear combinations of the weight and inputs.

Technical settings: For the calculations the python tools scikit-learn, keras, numpy and panda is used. Thereby the set of labels predicted for a sample must exactly match the corresponding set of labels in y_{true}. For the deep NN we consider tanh and softmax as activation functions, 50 epochs and a batch size of 64; for knn we use the Minkowsi metric, 3 nearest neighbors, for decision tree/random forest the Gini criterion is used, number of trees: 10 times the number of considered grid boxes and a depth of 5–10, where mostly 5 layers hold the best results. For the log regression the l2 penalty is considered.

Accuracy score: For all methods, we consider *sklearn .metrics.accuracy_score*, that computes the subset accuracy. Let $\hat{y}_i$ be the predicted value of the i-th sample and y_i the corresponding true value, then the fraction of correct predictions over the number of samples n is defined as

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}(\hat{y}_i = y_i) \tag{2.5}$$

with the indicator function $\mathbf{1}(x)$. See the scikit-learn manual for further information. From definition Eq. (2.5) it follows that the score is real valued and greater or equals to 0, and smaller or equals to 1.

3 Data

For the evaluation of the methods summarized in Section 2 to address research questions 1, 2 and 3 the COSMO-REA6 data set is used. The reanalysis data set is based on the non-hydrostatic, numerical weather prediction model COSMO of the German Weather Service (Deutscher Wetterdienst) with a continuous nudging scheme, see Bollmeyer et al. (2015). The COSMO-REA6 data set has a horizontal resolution of about 6 km and 40 vertical layers. We mainly consider the variables total precipitation and CAPE. The initial temporal resolution is one hour. The 3D wind, the temperature and the geopotential for level 21, which is about 600 hPa, are used to calculate the Dynamic State Index, see Eq. (3.3). The months July and August during the years 2013–2015 are analyzed. Spatially, we consider a box bounded by the latitudes 47.71°–54.74° N and the longitudes 2.50°–14.16° E. This domain contains parts of the Netherlands, Belgium, France, Germany, and parts of the Czech Republic. To compare the relation of the variables with respect to grid boxes of different sizes we average the original data. The spatial location of the domains are illustrated in Figure 1. See also the caption for a more precise explanation of the location of the domains.

3.1 Preprocessing 1: Calculating the input variables

Before the data are spatially averaged for the scale dependent analysis, the accumulated precipitation data has to be separated into a hourly time resolution. In the last part of our study we consider the Dynamic State Index (DSI) as additional predictor variable. This Index is calculated before further steps could be taken. The DSI is derived from the primitive equations and indicates atmospheric developments by unifying the information of the energetic and the vorticity state of the atmosphere (Névir, 2004). We denote with ρ the density, with $\boldsymbol{\xi}_a$ the absolute vorticity vector, $\boldsymbol{v}$ is the 3D wind vector, c_p the specific heat constant for dry air, T is the temperature and θ the potential temperature. Then, the DSI is defined as the Jacobi-determinant of the gradients of the potential vorticity

$$\Pi = \rho^{-1} \boldsymbol{\xi}_a \cdot \nabla\theta \,, \tag{3.1}$$

which gives the vortex-information, the gradient of the Bernoulli function

$$B = \frac{1}{2}\boldsymbol{v}^2 + c_p T + \phi \,, \tag{3.2}$$

which contains the kinetic energy, and the gradient of the potential temperature

$$\text{DSI} := \frac{\partial(\Theta, B, \Pi)}{\partial(a, b, c)} = \frac{1}{\rho} \frac{\partial(\Theta, B, \Pi)}{\partial(x, y, z)} = \frac{1}{\rho} (\nabla\theta \times \nabla B) \cdot \nabla\Pi \tag{3.3}$$

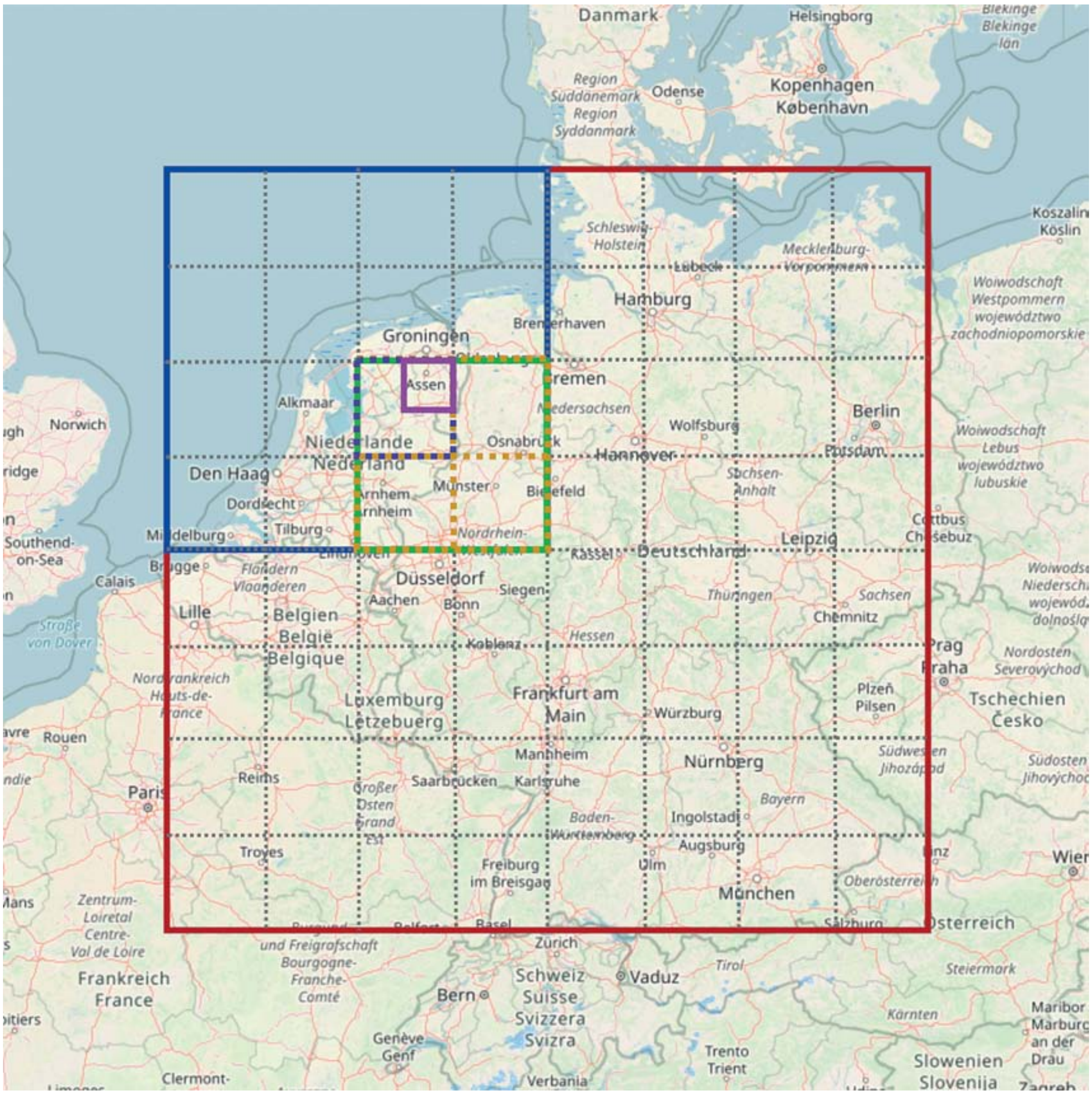

Figure 1: The large red box marks the large domain of 768^2 km^2. The blue box marks the region we consider for the data of the 384^2 km^2 domain, the green box labels the 192^2 km^2 domain, the dashed orange boxes are the boxes considered for the calculations with the resolution of 96^2 km^2, the same region is considered for the data in boxes of size 48^2 km^2 and 24^2 km^2. The dashed purple quarter of the yellow box is the domain for the 12^2 km^2 boxes and the pink box denotes the region containing the original 6^2 km^2 domains. In this purple domain fit in total 64 grid boxes of size 6^2 km^2.

with the Lagrangian mass coordinates a, b, c: $dm = da\,db\,dc = \rho\,dx\,dy\,dz$ and the density ρ.

The DSI is defined such that it is zero under adiabatic, inviscid, steady conditions (Névir, 2004). On the other hand, DSI signals unequal to zero indicate diabatic, viscous and non-steady states of the atmospheric flow field. Previous works have shown that DSI values unequal to zero are correlated to diabatic processes, especially precipitation (in a height of about 600 hPa), see e.g. Müller et al. (2018); Claussnitzer and Névir (2009). To differentiate more specific between moist air with and without phase changes and precipitating air, a complex hierarchy of DSI variants for these moist processes is developed by Hittmeir et al. (2021) and could be applied in future studies on the relation of CAPE and precipitation. Moreover, illustrating DSI fields show that the DSI is characterized by a dipole structure, which gives rise to the direction of motion, see e.g. Müller and Névir (2019). Additionally to the DSI we consider the so-called Thunderstorm Occurrence Parameter, short TOP, that combines DSI and CAPE:

$$TOP = |\,DSI|^{0.6} \cdot CAPE^{0.5}\,. \tag{3.4}$$

This parameter is introduced by Schartner et al. (2009).

3.2 Preprocessing 2: Averaging the input variables

The variables total precipitation, CAPE, DSI and TOP are temporally averaged over 3 hours, 4 hours, 6 hours, 12 hours and 24 hours. Furthermore, the data is spatially averaged such that we obtain data in grid boxes of sizes 6^2 km^2, 12^2 km^2, 24^2 km^2, 48^2 km^2 96^2 km^2, 192^2 km^2, 384^2 km^2 and 768^2 km^2. All averages are arithmetic means. The spatial domains are sketched in Figure 1.

3.3 Preprocessing 3: Classifying data in categories

For a first investigation of the relation of CAPE to precipitation for different temporal and spatial scales, we separate the total precipitation into the two categories *no rain* and *rain*, where only total precipitation intensities greater than the 75th percentile is considered. We remark that the sample sizes of both categories are equal. In case there are more events in the category *no rain* we took an additional random sample of the category *rain* to obtain equally sized samples. Thereby, we use the package *resample* from *sklearn.utils*.

Each of the CAPE, DSI, TOP values are divided into 10 categories, where the categories are the 10 %, 20 %, 30 %, ... percentiles, such that all categories have the same number of events.

3.4 Preprocessing 4: Training and test data

For all algorithms, we consider 70 percent of the data as training data and the remaining 30 percent of the data as test data set.

4 Results

In this section, we tackle all four research questions asked in the introduction. First, to gain an impression of the different variables we are evaluating, we show the spatial structures of CAPE, DSI and TOP and precipitation. Second, to answer the first two research questions, we compare the different machine learning algorithms summarized in Section 2. Third, we choose the algorithm with the highest score to calculate the relations of CAPE and precipitation across various temporal and spatial scales. To address the third research question, we take DSI and TOP as additional input variables into account and demonstrate their usefulness as parameters for finetuning. Finally, we show a theoretical relation of CAPE and precipitation intensity.

4.1 The spatial structure of CAPE, DSI, TOP and precipitation

One example of the spatial structure of the variables CAPE, DSI, TOP and precipitation is shown in Figure 2. The variables are represented for three hourly time steps. Comparing the CAPE field that is shown in the first row with the precipitation field depicted in the last row we recognize a time shift. Regarding CAPE, i.e. the *available* potential energy, we keep in mind that the energy does not has to be converted. But as higher the CAPE values, as higher is the probability that there will be convective activity, e.g. rain, whereas the DSI can be seen as a trigger parameter. Focusing the location of maximal CAPE values in the figure in the first row, first column, and comparing this figure with the figure of the precipitation intensity in the last row, last column, it can be recognized that the location of maximal CAPE values is reached two hours later by a convective cell with intense precipitation. We recognize a similar time shift of the DSI (second row) with the TOP index (third row), which can be explained as follows: The DSI is defined via gradients of the potential temperature, the Bernoulli function, containing the kinetic energy and the potential vorticity, see Eq. (3.3). Thus, the DSI identifies local changes, i.e. the developments of different processes, such as the *approaching* of storms before the rain falls out. The TOP index is defined as the product of CAPE and DSI, see Eq. (3.4). This variable combines CAPE and DSI and concentrates on the regions of strong convective activity, such as thunderstorms.

4.2 Using machine learning algorithms to analyze the relation of CAPE and precipitation

We will concentrate on only two precipitation cases: *rain* and *no rain*, where the class *rain* contains precipitation intensities equal and above the 75th percentile of all considered precipitation events. For example, for the original data with a horizontal resolution of 6 km and a one hour time resolution, a precipitation intensity of 0.7 mm/h is taken as threshold. The CAPE categories are classified into 10 classes. While we will investigate standard machine learning techniques, Polzin et al. (2022) investigate the relation of CAPE to the vertical velocity using an alternative approach called Direct Bayesian Model Reduction Gerber and Horenko (2017) using reanalysis data. In contrast, Gottwald et al. (2016) analyse data-driven stochastic models of tropical convection by using observations of the rain rate to build an entirely observation-based stochastic model.

4.2.1 Comparison of different machine learning methods

Regarding all outcomes of the methods logistic regression, perceptron, support vector machine (SVM), decision tree, random forest, k-nearest neighbor, and the simple deep neural network, over all, we find that the decision tree/random forest algorithms show the highest accuracy scores across all spatial and temporal scales. The deep neural network shows similar results. But since it takes more computational effort to calculate the deep neural network, we suggest the use of the decision tree/random forest algorithm. But we note, that every other algorithm did show good results for distinguished spatial-temporal scales.

In Table 1 the scores of the different algorithms are shown for two exemplary data sets. The second column is the score for the data averaged to domains of the size 96^2 km^2 and the 24 h mean is evaluated. The two precipitation categories are separated by a threshold of 0.091 mm, which seems to be a small number, but we recall that we take the arithmetic mean which leads to a small value of the 75th percentile of the considered data set. The averaging leads to a total number of 262 data

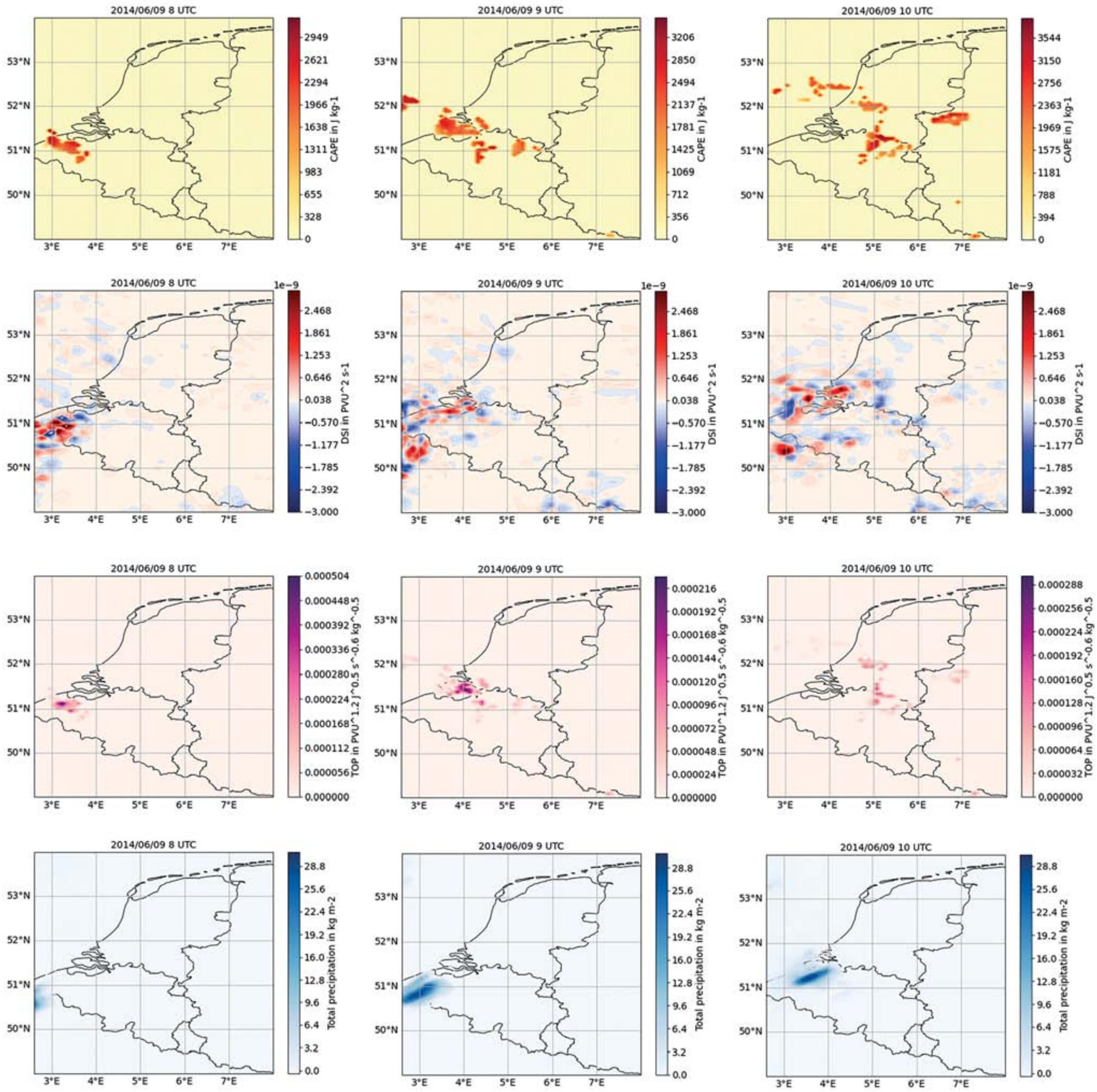

Figure 2: The variables CAPE (first row), DSI (second row), TOP (third row) and total precipitation (last hour accumulated; fourth row) are shown for the 2014/06/09 8 UTC (first column), 9 UTC (second column) and 10 UTC (third column). All three variables CAPE, DSI and TOP indicate the region of intense precipitation before it is raining.

points for this comparison. In this case, almost all methods show similar and good results. Only the k-nearest neighbour algorithm does not work as good. The third column shows the scores for the boxes of size 24^2 km^2. For this resolution we obtain in total 2828 data points. For this setting we consider precipitation categories with intensity smaller and greater than 0.143 mm per hour, and 10 CAPE categories. But there are examples, where the svm algorithm shows similar scores. For example, for grid boxes of the size 48 km x 12 km we obtain a score of 0.750 for the training as well as for the test data set. Applying the decision tree algorithm to this data leads to a score of 0.750 and 0.726 for the training and test data.

Regarding all results, we remark that we used the same technical settings for all scales, which can lead to over- and/or underfitting. Adapting these settings more precisely for every scale would optimize the results. Overall, the decision tree and random forest algorithms show the best scores. The classical logistic regression

Table 1: The second column is the score for the data averaged for regions of a horizontal resolution of 96 km, i.e. Germany is divided into 8 regions, and temporally the data is averaged to the 24 h mean. The third row shows the scores for the case that we divide Germany into smaller subdomains regarding a temporal mean of 24 hours. The depth for the decision tree and random forest algorithms is five.

	accuracy	accuracy
Method	hor. res 96 km	hor. res. 24 km
	24 h mean	24 h mean
Training accuracy log reg	0.705	0.715
Test accuracy log reg	0.696	0.720
Accuracy perceptron	0.696	0.482
Misclassified perceptron	24	440
Accuracy svm train	0.696	0.482
Accuracy svm test	0.696	0.757
Accuracy tree train	0.705	0.727
Accuracy tree test	0.696	0.757
Accuracy forest train	0.705	0.727
Accuracy forest test	0.696	0.757
Accuracy knn train	0.464	0.508
Accuracy knn test	0.582	0.482
Accuracy DNN train	0.705	0.715
Accuracy DNN test	0.697	0.720

shows almost as good scores as the tree algorithms. But there are cases, e.g. for the $6^2\,\text{km}^2$ domain, where the decision tree has higher scores than logistic regression algorithms (0.742 vs. 0.672). The scores of the deep neural network are in the same order as the scores of the decision trees for all scales. But these calculations need more time. The high scores for decision trees/random forest and the simple neural network can be explained by the fact that they capture nonlinear relations, i.e. they act as nonlinear mappings. In the following, we will stick to the relatively simple decision tree algorithm for further analysis of the temporal spatial dependencies of CAPE and precipitation in the next subsection.

4.2.2 Investigating temporal-spatial dependencies of CAPE and precipitation via decision trees

As explained in the previous paragraph, regarding all methods, the decision tree/random forest algorithms show the most consistent, highest accuracy scores across all scales.

Table 3 shows the accuracy scores of the spatial relation of CAPE to precipitation. As fix temporal resolution three hours are chosen, because of the time shift of the variables as exemplary shown in Figure 2. Regarding Table 3 the highest scores can be observed for about the same horizontal resolution, for the horizontal resolution of 24–96 km, the next higher order of resolution shows similar results. A reason might be the relation of the size of the convective event with the chosen time averaging of 3 hours, that fit together. Choosing a 24 hour average, we might get higher accuracy scores in the larger grid boxes.

Table 2: Accuracy of the method with best score for the relation of CAPE to precipitation via different temporal and spatial scales, each box: 1st row: accuracy of the training data, 2nd row: accuracy of the test data. The blue marked values are the output of the original model resolution with no averaging.

Resolution	1 h	3 h	4 h	6 h	12 h	24 h
768 km	0.656	0.697	0.753	0.696	0.750	0.769
	0.691	0.720	0.725	0.697	0.857	0.724
384 km	0.739	0.763	0.780	0.754	0.752	0.852
	0.721	0.716	0.752	0.781	0.765	0.820
192 km	0.725	0.763	0.727	0.705	0.727	0.796
	0.712	0.720	0.679	0.696	0.744	0.708
96 km	0.723	0.700	0.712	0.721	0.763	0.705
	0.723	0.745	0.725	0.674	0.720	0.696
48 km	0.735	0.727	0.734	0.737	0.743	0.721
	0.732	0.725	0.723	0.729	0.718	0.674
24 km	0.695	0.736	0.736	0.725	0.725	0.727
	0.692	0.722	0.728	0.731	0.719	0.757
12 km	0.690	0.681	0.684	0.723	0.720	0.732
	0.692	0.693	0.679	0.733	0.743	0.738
6 km	0.713	0.693	0.703	0.690	0.705	0.742
	0.715	0.693	0.690	0.702	0.691	0.724

The relation of CAPE to precipitation for different spatial and temporal scales calculated with the decision tree algorithm is presented in Table 2. We recognize scores mostly greater than 0.7 across all temporal and spatial scales. The 24 h mean shows the highest scores across all spatial scales. This might be explained as follows: the time shift of CAPE and precipitation, as discussed in the first paragraph of this section and illustrated in Figure 2, is 1–3 of hours and captured by the 24 h mean. We recall that CAPE indicates the *available* potential energy that will not necessarily be transformed. Detecting a CAPE signal, it can take 1–3 hours until we can measure precipitation. During this time, clouds might be moved. Dynamic processes explain the weaker scores for the one hourly data and small spatial resolutions. On the other hand this leads to higher scores for the 24 hour means as shown in the last column in Table 2.

We recall that we use the same technical configurations for all time and temporal scales, which can lead to over- and underfitting. Further technical adjustment would approve the results.

4.3 Considering the DSI and the TOP index as additional predictor variable

Due to the definition of CAPE there are cases with CAPE values greater than zero, but without precipitation. Therefore, to optimize the relation of CAPE and precipitation we suggest to take the Dynamic State Index (DSI) defined in Eq. (3.3) as additional input parameter into account. On the one hand, the DSI is zero for the adiabatic, inviscid basic state, such as persistent high pressure areas (Müller and Névir, 2019). On the

Table 3: Accuracy of the method with best score for different grid box sizes for the 3 h-averaged data X: CAPE, y: precipitation, the decision tree with 5 layers is used.

	768 km	384 km	192 km	96 km	48 km	24 km	12 km	6 km
768 km	0.710 0.707	0.492 0.588	0.601 0.545	0.589 0.593	0.579 0.570	0.612 0.612	0.587 0.582	0.621 0.624
384 km		0.763 0.716	0.640 0.566	0.603 0.649	0.588 0.591	0.596 0.594	0.598 0.585	0.592 0.580
192 km			0.763 0.720	0.705 0.739	0.720 0.722	0.708 0.724	0.578 0.569	0.591 0.581
96 km				0.700 0.745	0.723 0.738	0.715 0.738	0.556 0.552	0.573 0.568
48 km					0.727 0.725	0.719 0.736	0.556 0.544	0.569 0.567
24 km						0.722 0.736	0.549 0.529	0.564 0.552
12 km							0.708 0.721	0.512 0.518
6 km								0.693 0.693

Table 4: Accuracy of the method with best score for the relation of the two input variables CAPE and the DSI to precipitation via different temporal and spatial scales, each box: 1st row: accuracy of the training data, 2nd row: accuracy of the test data. Using the DSI as additional predictor increases the scores across all scales.

hor. resolution	1 h	3 h	4 h	6 h	12 h	24 h
768 km	0.746 0.722	0.767 0.720	0.802 0.762	0.743 0.697	0.789 0.750	0.831 0.793
384 km	0.775 0.783	0.781 0.813	0.829 0.780	0.835 0.750	0.829 0.804	0.889 0.869
192 km	0.747 0.741	0.796 0.685	0.750 0.705	0.732 0.759	0.788 0.767	0.815 0.667
96 km	0.739 0.739	0.754 0.752	0.747 0.741	0.752 0.720	0.732 0.747	0.732 0.747
48 km	0.749 0.743	0.740 0.729	0.749 0.730	0.756 0.746	0.752 0.720	0.752 0.720
24 km	0.714 0.714	0.744 0.736	0.748 0.745	0.737 0.748	0.747 0.729	0.747 0.729
12 km	0.714 0.714	0.716 0.735	0.722 0.711	0.749 0.740	0.742 0.723	0.752 0.741
6 km	0.730 0.730	0.726 0.726	0.741 0.734	0.738 0.737	0.744 0.711	0.762 0.749

other hand, non-zero valued DSI dipole structures can be used to indicate atmospheric developments such as hurricanes (Weber and Névir, 2008) or precipitation (Müller et al., 2018). Therefore, we propose the combination of CAPE and DSI as input variables to optimize the classification of the target variable precipitation. The results for the spatial-temporal relations are shown in Table 4 and the results of the spatial relations are summarized in Table 5. Indeed, taking the DSI additionally into account improves the scores across all temporal and spatial scales.

Moreover, we consider the TOP index defined in Eq. (3.4) as a further parameter, but the scores did not optimize the results compared to the DSI. This can be explained by the definition of TOP, which is designed to capture local extreme events such as intense thunderstorms. The TOP index is useful to identify thunderstorms and intense precipitation. But in this work, we only discuss two precipitation classes *rain* and *no rain* and do not further distinguish between precipitation events of different intensity. We propose the TOP index for further studies on precipitation extremes.

4.4 Theoretical relation of CAPE to precipitation

In order to answer research question 4, where we ask for a theoretical relationship of CAPE and precipitation we start with the frequently discussed relation of CAPE to the vertical velocity. The derivation is for example given by Holton (2004). We recall the theoretical relation of the vertical velocity and precipitation intensity introduced by Ertel (1933) and from this we shortly derive the theoretical relation of CAPE and precipitation.

The Convective Available Potential Energy, short CAPE, is given by the integral from the level of free convection (LFC) to the equilibrium level (ET), as defined in Eq. (1.2). Reformulating the vertical momentum equation leads to the following quadric relation of the maximal vertical velocity and CAPE:

$$\frac{v_z^2}{2} = \text{CAPE} \quad \Longleftrightarrow \quad v_z = \sqrt{2\cdot\text{CAPE}}\,. \tag{4.1}$$

Table 5: Accuracy of the method with best score for different grid box sizes for the 3 h-averaged data X: CAPE, DSI, y: precipitation, the results of the random forest algorithm are shown providing the highest scores.

	768 km	384 km	192 km	96 km	48 km	24 km	12 km	6 km
768 km	0.765 0.729	0.714 0.628	0.742 0.671	0.716 0.723	0.691 0.673	0.720 0.736	0.740 0.736	0.649 0.656
384 km		0.829 0.742	0.733 0.580	0.686 0.626	0.666 0.665	0.668 0.670	0.716 0.715	0.722 0.704
192 km			0.826 0.636	0.767 0.752	0.773 0.766	0.759 0.764	0.683 0.993	0.722 0.723
96 km				0.767 0.735	0.757 0.748	0.740 0.744	0.597 0.585	0.652 0.639
48 km					0.760 0.745	0.736 0.740	0.573 0.572	0.605 0.591
24 km						0.744 0.750	0.561 0.539	0.591 0.577
12 km							0.730 0.743	0.543 0.532
6 km								0.729 0.724

The relation of the vertical velocity and precipitation has been analysed numerically and statistically, see e.g. Pendergrass and Gerber (2016), Weijenborg et al. (2017), or Müller et al. (2020), but it is only rarely discussed theoretically. Almost ninety years ego, Ertel (1933) suggested the following relation between the vertical velocity v_z and the precipitation intensity I:

$$v_z = \frac{39.2}{\log\left(\frac{\theta_H}{\theta_h}\right)} \frac{I}{\overline{p}} . \tag{4.2}$$

We follow the dimensions used in Ertel (1933), where the precipitation intensity I is given in mm/(60 min). Moreover, $\overline{p}$ denotes the arithmetic mean of the pressure between the heights *LFC* and *ET*. Now, we combine the quadric relation of CAPE and the vertical velocity Eq. (4.1) with Ertels relation of the vertical velocity and the precipitation intensity Eq. (4.2) and obtain the quadratic relation of CAPE and the precipitation intensity:

$$\begin{aligned} \mathrm{CAPE} = \frac{v_z^2}{2} &= \frac{1}{2}\left(\frac{39.2}{\log\left(\frac{\theta_H}{\theta_h}\right)} \frac{I}{\overline{p}}\right)^2 \\ &= 768.32\left(\frac{I}{\log\left(\frac{\theta_{ET}}{\theta_{\mathrm{LFC}}}\right)\overline{p}}\right)^2 . \end{aligned} \tag{4.3}$$

It follows that:

$$\mathrm{CAPE} \propto I^2 . \tag{4.4}$$

Therefore, we theoretically assert that there *is* a relation of CAPE and precipitation. But we leave the verification of Eq. (4.3) to future work. In general, since the computational effort increases exponentially, we think machine learning approaches should be considered not to substitute, but to complement the results of the theoretical equations of motions of atmospheric dynamics.

5 Conclusion

In this work, we use machine learning algorithms to investigate the scale dependent relation of the predictor variable CAPE and the predictand variable precipitation taking additionally DSI and the TOP as predictor variables into account. The goal was to tackle the following four research questions: *(1) Can we use todays increasing possibilities of the applicability of machine learning algorithms to investigate CAPE and precipitation across various temporal and spatial scales? (2) Which algorithms provide the best relations of CAPE and convective activity? (3) Which benefits provide the Dynamic State Index (DSI) and the Thunderstorm Occurrence Parameter (TOP) as additional input variable? and (4) Is there a theory of the relation of CAPE with precipitation?* To answer these question, the COSMO-REA6 data set with an initial horizontal resolution of 6 km and a spatial resolution of one hour are used. In order to obtain the lower resolutions to analyze relations of the variables across different spatial and temporal scales, the data is filtered. We finally analyze data of a temporal resolution of 24 hours up to one hour. Furthermore, we regard different spatial scales with horizontal resolutions of 6 km, 12 km, 24 km, 96 km, 48 km, 96 km, 192 km, 384 km and 768 km. For each resolution we consider two precipitation classes (rain/no rain) and 10 classes of each of the predictor variables CAPE and DSI.

To answer the first two questions, we start with a comparison of the accuracy scores of the different methods logistic regression, perceptron, support vector machine, decision tree and random forest, k-nearest neighbor and a simple deep neural network. Overall, we find that the decision tree algorithm and random forest provide the best scores. The deep neural network shows very similar scores. Both models are nonlinear classifiers. These findings confirm the results of Grazzini et al. (2020), who investigate a methodology for identification and systematic classification of extreme precipitation events over northern central Italy. They find that the random forest classifier turn out to be decisive in finding an optimal classification and for neglecting non-useful predictors.

As we have shown in Section 4.4 theoretically, the relation of CAPE and precipitation intensity is quadratic, i.e. nonlinear. This explains the over all good results of the decision tree algorithm and the deep neural network. Because of the additional computer effort of deep neural networks, we concentrate on the decision tree algorithm for the analysis of the scale dependence relation. Overall, we find scores about 0.7 across all scales. We did use the same technical configurations for all temporal and spatial resolutions. Therefore, we think that specific technical finetuning would optimize the results. In this first study, we only consider two precipitation classes and ten CAPE categories. It is wishful to increase the number of categories for a more precise forecast of precipitation intensities.

To tackle the third research question, we take the Dynamic State Index (DSI) as additional input variable into account. This leads to a slightly increase of the accuracy scores. The DSI unifies local changes of the kinetic energy (via the Bernoulli function), of the potential temperature and of the vorticity into one scalar. This means that the DSI indicates different processes. While CAPE provides information about the convective *available* potential energy, the DSI indicates regions, where the release takes place.

Therefore, the DSI seems to be a suitable variable that should be additionally taken into account to optimize the relation of CAPE and precipitation. For further optimization, different scale-dependent DSI variants could be taken into account, such as the DSI variants for moist processes that are recently introduced by Hittmeir et al. (2021). We also regard the TOP index as predictor variable, which is especially useful for the identification of intense precipitation events. Since we considered here only two precipitation classes, the TOP index did not improve our results. But it seems to be an interesting variable for studies on precipitation extremes.

The classification of precipitation into two classes leads to scale independent results. Considering more classes of precipitation intensity might lead to scale dependent scores: lower scores for high resolutional data and higher scores for lower resolutional data, which will be analyzed in future studies. Here, the accuracy score that counts the number of correctly classified data instances over the total number of data instances is evaluated to compare relations via different models across different scales. For a specific analysis at a certain scale further scores could be taken into account.

In order to answer the fourth initially asked research question, we show a theoretical relation of CAPE and precipitation. This should be evaluated in future work in order to e.g. prognose the intensity of precipitation in more detail.

We conclude that we suggest the use of machine learning algorithms, especially decision trees, to complement parameterization schemes. Decision trees are simple nonlinear networks. Of course, deep learning networks are also nonlinear classifiers. But neural networks need to be designed more complex for a higher score, which would take more computational time. In general we think that machine learning can not and should not be applied to substitute the atmospheric equations of motions. But the increasing computer effort and the progress in the enhancements of the coding of machine learning algorithms makes it almost naturally to use machine learning algorithms to *complement* the results gained by the equations of motions. Especially, the machine learning approach is helpful for learning the evolution of complex systems and processes such as convection.

Acknowledgements

We thank the anonymous reviewer for the helpful comments and suggestions. We acknowledge support by the German Research Foundation for their support within the framework of CRC 1114 'Scaling Cascades in Complex Systems', project A01 and the Open Access Publication Fund of TU Berlin.

References

Adams, D.K., E.P. Souza, 2009: Cape and convective events in the southwest during the north american monsoon. – Mon. Wea. Rev. **1**, 83–98, DOI: 10.1175/2008MWR2502.1.

Bollmeyer, C., J. Keller, C. Ohlwein, S. Wahl, S. Crewell, P. Friederichs, A. Hense, J. Keune, S. Kneifel, I. Pscheidt, others, 2015: Towards a high-resolution regional reanalysis for the european cordex domain. – Quart. J. Roy. Meteor. Soc. **686**, 1–15, DOI: 10.1002/qj.2486.

Claussnitzer, A., P. Névir, 2009: Analysis of quantitative precipitation forecasts using the dynamic state index. – Atmos. Res. **4**, 694–703, DOI: 10.1016/j.atmosres.2009.08.013.

Ertel, H., 1933: Die vertikale Luftbewegung bei Starkregen. – Meteorol. Z. **2**, 149–152.

Gerber, S., I. Horenko, 2017: Toward a direct and scalable identification of reduced models for categorical processes. – Proc. Natl. Acad. Sci. **19**, 4863–4868, DOI: 10.1073/pnas.1612619114.

Gottwald, G.A., K. Peters, L. Davies, 2016: A data-driven method for the stochastic parametrisation of subgrid-scale tropical convective area fraction. – Quart. J. Roy. Meteor. **694**, 349–359, DOI: 10.1002/qj.2655.

Grazzini, F., G.C. Craig, C. Keil, G. Antolini, V. Pavan, 2020: Extreme precipitation events over northern italy. part i: A systematic classification with machine-learning techniques. – Quart. J. Roy. Meteor. **726**, 69–85, DOI: 10.1002/qj.3635.

Hittmeir, S., R. Klein, A. Müller, P. Nèvir, 2021: The Dynamic State Index with moisture and phase changes. – J. Math. Phys. **62**, 1231091, DOI: 10.1063/5.0053751.

Holton, J.R., 2004: An Introduction to Dynamic Meteorology, volume 4. – Academic Press.

Khouider, B., 2019: Models for tropical climate dynamics: waves, clouds, and precipitation, volume 3. – Springer.

Markowski, P., Y. Richardson, 2011: Mesoscale meteorology in midlatitudes, volume 2. – John Wiley & Sons.

Müller, A., P. Névir, 2019: Using the concept of the dynamic state index for a scale-dependent analysis of atmospheric blocking. – Meteorol. Z. **28**, 487–498, DOI: 10.1127/metz/2019/0963.

Müller, A., P. Névir, R. Klein, 2018: Scale dependent analytical investigation of the dynamic state index concerning the quasi-geostrophic theory. – Math. Climate Wea. Forecast **1**, 1–22, DOI: 10.1515/mcwf-2018-0001.

Müller, A., B. Niedrich, P. Névir, 2020: Three-dimensional potential vorticity structures for extreme precipitation events on the convective scale. – Tellus A **1**, 1–20, DOI: 10.1080/16000870.2020.1811535.

National Weather Service, 2023: Severe Weather Topics. – Published online, https://www.weather.gov accessed: 2023-02-2.

Névir, P., 2004: Ertel's vorticity theorems, the particle relabelling symmetry and the energy-vorticity theory of fluid mechanics. – Meteorol. Z. **6**, 485–498, DOI: 10.1127/0941-2948/2004/0013-0485.

Pendergrass, A.G., E.P. Gerber, 2016: The rain is askew: Two idealized models relating vertical velocity and precipitation distributions in a warming world. – J. Climate **18**, 6445–6462, DOI: 10.1175/JCLI-D-16-0097.1.

Polzin, R., A. Müller, H. Rust, P. Névir, P. Koltai, 2022: Direct bayesian model reduction of smaller scale convective activity conditioned on large-scale dynamics. – Nonlin. Process. Geophys. **29**, 37–52, DOI: 10.5194/npg-29-37-2022.

Ramezani Ziarani, M., B. Bookhagen, T. Schmidt, J. Wickert, A. de la Torre, R. Hierro, 2019: Using convective available potential energy (cape) and dew-point temperature to characterize rainfall-extreme events in the south-central andes. – Atmosphere **7**, 379, DOI: 10.3390/atmos10070379.

Raschka, S., V. Mirjalili, 2017: Python machine learning: Machine learning and deep learning with python. – Scikit-Learn, and TensorFlow.

Rennó, N.O., A.P. Ingersoll, 1996: Natural convection as a heat engine: A theory for cape. – J. Atmos. Sci. **4**, 572–585, DOI: 10.1175/1520-0469(1996)053<0572:NCAAHE>2.0.CO;2.

Riemann-Campe, K., K. Fraedrich, F. Lunkeit, 2009: Global climatology of convective available potential energy (cape) and convective inhibition (cin) in era-40 reanalysis. – Atmos. Res. **1–3**, 534–545, DOI: 10.1016/j.atmosres.2008.09.037.

Rybka, H., U. Burkhardt, M. Köhler, I. Arka, L. Bugliaro, U. Görsdorf, Á. Horváth, C.I. Meyer, J. Reichardt, A. Seifert, J. Strandgren, 2020: The behavior of high-cape (convective available potential energy) summer convection in large-domain large-eddy simulations with icon. – Atmos. Chem. Phys. **6**, 4285–4318, DOI: 10.5194/acp-2020-635.

Schartner, T., P. Névir, G. Leckebusch, U. Ulbrich, 2009: Analysis of thunderstorms with the dynamic state index (dsi) in a limited area high resolution model. – In: 5th European Conference on Severe Storms, 12–16.

Weber, T., P. Névir, 2008: Storm tracks and cyclone development using the theoretical concept of the dynamic state index (dsi). – Tellus A **1**, 1–10, DOI: 10.1111/j.1600-0870.2007.00272.x.

Weijenborg, C., J. Chagnon, P. Friederichs, S. Gray, A. Hense, 2017: Coherent evolution of potential vorticity anomalies associated with deep moist convection. – Quart. J. Roy. Meteor. **704**, 1254–1267.

Weisman, M.L., J.B. Klemp, 1982: The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. – Mon. Wea. Rev. **6**, 504–520, DOI: 10.1175/1520-0493(1982)110<0504:TDONSC>2.0.CO;2.

Williams, E., N. Renno, 1993: An analysis of the conditional instability of the tropical atmosphere. – Mon. Wea. Rev. **1**, 21–36, DOI: 10.1175/1520-0493(1993)121<0021:AAOTCI>2.0.CO;2.