JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2018MS001561

Evaluation of Machine Learning Classifiers for Predicting Deep Convection

Peter Ukkonen[1] and Antti Mäkelä[2]

[1]Danish Meteorological Institute, Copenhagen, Denmark, [2]Finnish Meteorological Institute, Helsinki, Finland

Key Points:
- Machine learning models are able to predict thunderstorm occurrence with much greater skill than traditional parameters
- The best parameters for predicting thunderstorms vary by region, but neural networks perform well across different climates if trained on them
- Only by using machine learning can the diurnal cycle and climatology of thunderstorm occurrence be faithfully reproduced from reanalysis data

Correspondence to:
P. Ukkonen
peter.ukkonen@nbi.ku.dk

Citation:
Ukkonen, P., & Mäkelä, A. (2019). Evaluation of machine learning classifiers for predicting deep convection. *Journal of Advances in Modeling Earth Systems*, *11*, 1784–1802. https://doi.org/10.1029/2018MS001561

Received 13 NOV 2018
Accepted 28 APR 2019
Accepted article online 9 MAY 2019
Published online 21 JUN 2019

Abstract The realistic representation of convection in atmospheric models is paramount for skillful predictions of hazardous weather as well as climate, yet climate models especially suffer from large uncertainties in the parameterization of clouds and convection. In this work, we examine the use of machine learning (ML) to predict the occurrence of deep convection from a state-of-the-art atmospheric reanalysis (ERA5). Logistic regression, random forests, gradient-boosted decision trees, and deep neural networks were trained with lightning data to predict thunderstorm occurrence (TO) in Central and Northern Europe (2012–2017) and in Sri Lanka (2016–2017). Up to 40 input variables were used, representing, for example, instability, humidity, and inhibition. Feature importances derived for the various models emphasize the high importance of conditional instability for deep convection in Europe, while in Sri Lanka, TO is more strongly regulated by humidity. The Precision-Recall curve indicates more than a twofold improvement in skill over convective available potential energy for short-term (0–45 min) predictions of TO in Europe by using neural networks or gradient-boosted decision tree and a larger improvement in the tropical domain. The diurnal cycle of deep convection is closely reproduced, suggesting that ML could be used to trigger convection in climate models. Finally, a strong relationship was found between area-mean monthly TO and ML predictions, with correlation coefficients exceeding 0.94 in all domains. Convective available potential energy has a similar level of correlation with monthly thunderstorm activity only in Northern Europe. The results encourage the use of reanalyses and ML to study climate trends in convective storms.

1. Introduction

Hazards associated with convective weather (e.g., lightning, large hail, tornadoes, downbursts, and heavy precipitation) cause several billions of euros worth of damage in Europe on a yearly basis (Dotzek et al., 2009). Besides the high direct impact of convective weather on society, convection is also an important role in Earth's climate by producing clouds which affect the radiation budget and transport heat and moisture in the vertical. It is therefore paramount to represent convective processes accurately in atmospheric models; unfortunately, this also happens to be a difficult task. Global and regional climate models (GCMs and RCMs) are known to suffer from many unrealistic aspects in the simulated convection. Systematic biases have been found in the intensity distribution (Kyselý et al., 2015) and diurnal cycle of precipitation (Dirmeyer et al., 2011), the vertical structure of convective heating and moistening (Herman & Kuang, 2013), and cloud cover (Cesana & Waliser, 2016). Such errors are known to have a major impact on the overall skill of GCM simulations (Sherwood et al., 2014).

Recently, it has become possible to circumvent many of these issues by running RCMs on resolutions where deep convection is at least partially resolved, whereby the deep convection scheme can be turned off. This is a significant development which has led to clear improvements in, for example, the statistics of simulated convective precipitation, the shallow-to-deep convection transition, and in convective aggregation (Ban et al., 2014; Kendon et al., 2017; Prein et al., 2015). A recent example is Knist et al. (2018), who managed to run three 12-year convection-permitting simulations over central Europe. However, the computational cost of such simulations is enormous, and it will take many years before larger domains and longer time periods can be used.

Studies on climate change impacts on convective weather are also hampered by deficiencies in convective-parameterizing models. Such studies have generally focused on severe convective storms in the United States and used the combination of convective available potential energy (CAPE) and wind shear as a proxy for these storms (Brooks, 2013). Modeling studies have pointed toward an increased frequency of

convective storms in the future as a result of robust increases in CAPE across both the United States and Europe (Púčik et al., 2017). However, there is still much uncertainty regarding changes in wind shear in many regions, as well as storm initiation (Allen, 2018). One alternative is using the observational record; unfortunately, systematic records of convective weather tend to be spatially or temporally limited and/or inhomogeneous. In Finland, records exist since 1887 as manual thunderstorm day observations, since 1960 using automatic lightning flash counters, and since the early 1980s as lightning location system observations (Mäkelä et al., 2014b). Such changes in the observing system, seen also in other countries, make the record unreliable for the estimation of climate trends (Brooks, 2013).

In light of these problems, what is the way forward? The convective parameterization issue will stay relevant for years to come, and progress thus far has been slow despite considerable efforts using different theoretical frameworks (Yano et al., 2015). An alternative approach, which could offer a breakthrough, is to learn the complex nonlinear relationships governing subgrid convection directly from observational and numerical modeling data. This can be done using machine learning (ML) algorithms such as neural networks (NNs). Papers published while this work was finalized suggest a sudden jump in interest in the use of ML to parameterize moist convection and other subgrid processes in climate models (Brenowitz & Bretherton, 2018; Gentine et al., 2018; Gorman & Dwyer, 2018; Rasp et al., 2018). One approach is to train a ML parameterization using output from a cloud-resolving model (CRM) embedded in a climate model, also known as a superparameterization. Gentine et al. (2018) demonstrated this method using idealized aquaplanet simulations and obtained promising results in terms of skillful predictions of convective heating, moistening, and radiative features of the superparameterization. In a similar study by Rasp et al. (2018), multiyear GCM simulations incorporating NN-based parameterizations were shown to closely reproduce the mean climate of CRM simulations, as well as key aspects of variability. However, some loss of variability was seen in both of these studies, resulting from NNs being inherently deterministic. A further drawback of machine-learned parameterizations is that physical properties on various scales may not be conserved unless explicitly accounted for. Brenowitz and Bretherton (2018) was able to develop a numerically stable NN parameterization by minimizing the prediction error over multiple time steps rather than a single one, but, for example, column-integrated moist static energy was not conserved.

We propose that a promising application for ML may lie in the "trigger function" which activates the deep convective scheme in GCMs. Triggering convection at the right time and place is important for the realistic simulation of atmospheric variability, yet existing trigger functions tend to be simplistic by design, using near-arbitrary thresholds and ignoring important processes. As a result, convection is often triggered too easily (Suhas & Zhang, 2014). Inadequate triggering criteria have been linked to unrealistic simulations of, for example, the diurnal cycle of convection (Xie, 2004), the Madden-Julian Oscillation (Lin et al., 2008), and the intertropical convergence zone (Liu et al., 2009b). The nonlinearity and possible threshold behavior (Houston & Niyogi, 2007; Yano et al., 2012) of deep convective initiation makes this problem a good candidate for machine-learned classification models, which predict probabilities and therefore offer stochasticity. An ML-based stochastic convective trigger could, for example, be used alongside existing convective schemes in GCMs.

In this paper, we use state-of-the-art reanalysis and lightning data to evaluate the skill of ML models to predict the occurrence of deep convection in different climates. This serves as a data-rich platform to test the hypothesis that ML can greatly improve the prediction of deep convection. Lightning activity can be monitored accurately around the globe using remote sensing (Ávila et al., 2010) and is closely related to the vertical velocity and cloud depth of convective clouds (Williams, 2001). Reanalyses, meanwhile, combine millions of historical observations with modern data assimilation and numerical modeling to give a "best guess" of the past atmospheric state in a consistent manner. Reanalyses, NNs, and lightning data were previously combined in Ukkonen et al. (2017), where categorical thunderstorm forecasts for Finland using shallow NNs had substantially higher skill compared to any single predictor.

We train logistic regression, decision tree ensembles, and deep NNs to predict the occurrence of thunderstorms from parameters related to mainly instability, inhibition, and moisture. Feature scores are used to explore factors regulating thunderstorm occurrence (TO) in the tropics and middle/high latitudes. The clas-

sifiers are evaluated in terms of prediction of individual events, the diurnal cycle, and the correlation with observed thunderstorms on larger scales.

Section 2 describes the data, preprocessing, ML algorithms, and experiment workflow. Model training and optimization is detailed in section 3. This is followed by evaluation (section 4) and a discussion on the environmental factors governing deep convection in different climates and on different scales (section 5). Finally, we offer some concluding remarks and discuss the implications of our findings (section 6).

2. Data and Methodology

2.1. Data

Reanalysis and lightning data from two European domains (Northern and Central Europe) with high-quality lightning location networks were obtained for the convectively active season (May–August) spanning 2012–2017. Data were also acquired for a tropical domain (Sri Lanka, January–December 2016–2017) in order to study differences between tropical and middle- to high-latitude convection and the potential for developing a global scheme. The domains are depicted in Figure 5.

2.1.1. ERA5

ERA5 is the fifth generation global reanalysis product by European Centre for Medium-range Weather Forecasts (ECMWF), set to replace ERA-Interim and cover the period from 1950 to present (Hersbach & Dee, 2016). ERA5 features a number of changes from ERA-Interim, with a key improvement being a much higher resolution: hourly analysis fields are available at a horizontal grid spacing of 31 km on 137 vertical levels. This should be sufficient to resolve mesoscale structures important for convective initiation. The Earth system model used in ERA5 is the Integrated Forecast (IFS) Cycle 41r2, which in comparison with the older IFS version 31r2 represents many developments in physics parameterizations and data assimilation methods. For example, ERA5 uses a variational bias scheme not only for satellite radiances as in ERA-Interim but also for ozone, aircraft, and surface pressure data. ERA5 data were acquired on longitude-latitude grids using different longitudinal grid increments for each of the three domains in order to have a roughly 28-km horizontal grid spacing everywhere. Training data were obtained at 2-hourly time step, resulting in nearly 28 million pseudo-soundings for Europe.

We briefly evaluated the quality of ERA5 pseudo-soundings by comparing values of CAPE derived from ERA5 and radiosonde data. The 00 and 12 UTC soundings from five radiosonde stations in Northern Europe (N-EUR) and 12 stations in Central Europe (C-EUR) were obtained for 2012–2017, most of them located in Sweden and Germany, respectively. The linear correlation between Mixed Layer CAPE derived from ERA5 pseudo-soundings and radiosonde measurements was 0.78 in C-EUR and 0.74 in N-EUR. The agreement is fairly good given that the stations in N-EUR (C-EUR) were located on average 9.5 (8.3) km from the reanalysis grid points they were compared to. Furthermore, the correlation between observed TO (section 2.2.1) and CAPE was similar for ERA5 and radiosonde data in C-EUR (r = 0.21–0.22). In N-EUR, sounding CAPE did have a stronger correlation with observed thunderstorms than ERA5-CAPE (0.28 and 0.21, respectively), which could imply a lower quality of ERA5 soundings in this region. The predecessor of ERA5, ERA-Interim, was compared extensively to sounding data in Europe by Taszarek et al. (2018). While boundary layer moisture and midtropospheric lapse rates were in general described very well by ERA-Interim, convective parameters such as most unstable CAPE and mixed-layer CIN still exhibited large relative errors in a Central European domain (35.2% and 33.9%, respectively).

2.1.2. Lightning Data

Lightning observations were used to assess the occurrence of deep, moist convection. The advantages of lightning observations are a very high detection efficiency and spatial accuracy, as well as a large spatial coverage. In the European domain, data from two lightning location networks were obtained. The Nordic Lightning Information System (Mäkelä et al., 2014b), which covers the Nordic and Baltic countries in a combined network since 2002, is used for the Northern European domain. For the Central/West European domain, data were received from the EUropean Cooperation for LIghtning Detection network (Schulz, 2005), which is a collaboration among many national lightning networks in Europe and also encompasses the Nordic Lightning Information System network. In Sri Lanka, we have used data from the Global Lightning Dataset 360 of Vaisala that has been noted to have relatively homogeneous global performance (e.g., Pohjola & Mäkelä, 2013) and has been used in previous studies in Asia (Mäkelä et al., 2014a). Because our primary interest is on the yes/no occurrence of lightning, we have considered here all the located lightning events, that is, either cloud-to-ground (CG) or intracloud (IC) flashes in our analysis. Differences between

the data sets regarding the detection efficiency of IC flashes should not have a significant impact on the results given that, for example, a majority of midlatitude continental convective systems are associated with at least one CG flash (MacGorman et al., 2011).

2.2. Preprocessing Data

2.2.1. Predictand

To develop a classification model for predicting TO, class labels first need to be generated by associating lightning observations with sounding data by some spatiotemporal criteria. The choice of criteria is nontrivial and may have a large impact on results. In Ukkonen et al. (2017), ERA-Interim pseudo-soundings were labeled as "thundery" if at least one lightning flash was observed within the 0.75° × 0.75° grid box (2,100 − 3,100 km^2) during the 6-hr period after the analysis time. Other proximity criteria used in literature include within a 0.75° × 0.75° grid box and 0–2 hr of pseudo-sounding (Westermayer et al., 2017) and within 125 km and −2–6 hr of sounding (Taszarek et al., 2017). The former of this studies selected the temporal criteria based on 2 hr being the typical timescale for a thunderstorm to advect across a grid box assuming a typical wind of 10–15 m/s. We follow the same reasoning, which leads to a roughly 45-min time window for typical storms to leave the higher-resolution grid cells. Samples were consequently assigned to the thundery class if at least one CG or IC flash was observed within the 28 × 28-km grid cells within 45 min after the analysis. This strict criteria should lead to pseudo-soundings being highly representative of preconvective environments but results in a low ratio of thunderstorm events to nonevents.

2.2.2. Predictors

Model level data were acquired at high vertical resolution in order to resolve thermodynamically important structures such as capping inversions. Temperature and specific humidity were obtained on 42 vertical levels and wind fields on eight levels, emphasizing the low-to-middle troposphere in both cases. While it is possible to predict thunderstorms directly from vertical profiles, in our case this would entail over 100 inputs and would likely require deep NNs capable of transforming the temperature and humidity data into higher level representations associated with, for example, buoyancy and saturation. An alternative is to manually extract physically meaningful features from the vertical columns, for example, variables such as CAPE which are known to be important for convection and thereby reduce the dimensionality of the data. This process of leveraging domain specific knowledge and human insight is known as feature engineering. We use this approach as it enables the use of simpler models and facilitates model interpretation.

Unfortunately, this results in the difficult and time-consuming exercise of transforming soundings into a set of convective predictors which is not overly large and redundant and yet does not omit important information. We decided to first generate a large number of convective parameters from model level data. To do this in a flexible and computationally efficient manner, we wrote a Julia package (https://github.com/peterukk/ConvectiveIndices.jl) for calculating convective indices and thermodynamic variables from atmospheric profiles. Julia is a modern high-level, high-performance dynamic programming language designed for numerical computing. For computational reasons, parameters such as CAPE and Lifted Index are by default calculated using a θ_e formulation

$$CAPE = g \int_{z_{LFC}}^{z_{EL}} \left(\frac{\theta_{e,i} - \bar{\theta}_{es}}{\bar{\theta}_{es}} \right) \mathrm{d}z \tag{1}$$

$$LI_i = \theta_{es,500} - \theta_{e,i} \tag{2}$$

where g is the acceleration of gravity, θ_e is the equivalent potential temperature, which is conserved during a pseudo-adiabatic ascent, the subscripts i and 500 refer to the initial parcel level and 500-hPa level, $\bar{\theta}_{es}$ is the environmental saturated equivalent potential temperature, z_{LFC} is the level of the free convection, and z_{EL} is the equilibrium level. LFC and EL are here taken as the lowest level where the term inside the parenthesis is positive and the highest level where the term is negative, respectively. Furthermore, only positive contributions to the integral are included. These parameters were found to perform just as well as thunderstorm predictors as the traditional formulations of CAPE and Lifted Index (see, e.g., Ukkonen et al., 2017).

Having calculated numerous convective parameters, candidate predictors were evaluated mainly by using thunderstorm probability tables (Ukkonen et al., 2017). These tables depict the binned empirical thunderstorm probability as a function of two parameters. A clearly changing thunderstorm probability as a function

Table 1
List of Input Variables Used in the Study

Acronym	Description
blh	Boundary layer height
$CAPE_i$	Convective available potential energy using the θ_e approximation (equation (1)). Parcels i used in this study are the surface parcel (SFC), a mixed-layer parcel based on the mean conditions in the lowest 50 hPa (ML), and the most unstable parcel in the lowest 350 hPa (MU), which is furthermore mixed over a 50-hPa depth
$CAPECIN_i$	$CAPECIN_i = g\int_{z_{LCL_i}}^{z_{LCL_i-250hPa}} \left(\frac{\theta_{e,i}-\bar{\theta}_{es}}{\bar{\theta}_{es}}\right) dz$, where i refers to the parcel (MU or ML) and both negative (CIN) and positive contributions (CAPE) are included, and the LCL is the lifted condensation level
$CIN_{LCL,i}$	$CIN_{LCL,i} = g\int_{z_i}^{z_{LCL_i}} \left(\frac{\theta_{e,i}-\bar{\theta}_{es}}{\bar{\theta}_{es}}\right) dz$, where i refers to the parcel (MU or ML) or parcel level
d2m	2-m dew point temperature
DLS	Bulk wind shear between 10-m AGL (above ground level) and model level 94 (roughly 6-km AGL)
hcc	High cloud cover
ie	Instantaneous moisture flux
ishf	Instantaneous surface sensible heat flux
lcc	Low clover cover
LI_i	Lifted Index using the parcel i (MU, ML, or SFC). See equation (2)
LLS	Bulk wind shear between 10-m AGL and model level 118 (roughly 1-km AGL)
mcc	Medium clover cover
$MRH_{300–600}$	Mean relative humidity (%) between 300 and 600 hPa
$MRH_{600–800}$	Mean relative humidity (%) between 600 and 800 hPa
$MRH_{ALCL,i}$	Mean relative humidity (%) between the LCL and 250 above the LCL, using the parcel i (MU or ML)
msl	Mean sea level pressure
MLS	Bulk wind shear between 10-m AGL and model level 105 (roughly 3-km AGL)
ULS	Bulk wind shear between model level 105 (roughly 3-km AGL) and model level 94 (roughly 6-km AGL)
sp	Surface pressure
soiltemp	Soil temperature level 1
soilwater	Volumetric soil water layer 1
$pLCL_{MU}$	Lifted condensation level (hPa) of the most unstable parcel
t2m	2-m temperature
tciw	Total column cloud ice water
tcw	Total column water
tcwv	Total column water vapor
V_{mid}	Mean of the horizontal wind speeds at model level 105 (roughly 3-km AGL) and 94 (roughly 6-km AGL)
$VIHMC_{low}$	Vertically integrated horizontal mass convergence in the lowest 300 hPa
vimfc	Vertical integral of divergence of moisture flux
vimt	Vertical integral of mass tendency
vidfw	Vertical integral of divergence of cloud frozen water flux
w_{mid}	Mean of the vertical velocities at model level 105 (roughly 3-km AGL) and 94 (roughly 6-km AGL)
zBCL	Height of the buoyant condensation level (BCL), which quantifies atmospheric preconditioning to surface-triggered convection (Tawfik & Dirmeyer, 2014). Calculated using equation 1 in Tawfik et al. (2017)

Note. Variables which are abbreviated using only lowercase letters are ECMWF surface level parameters (the acronym may have been changed). The remaining variables have been calculated from model level data.

of a candidate predictor for constant values of a baseline variable such as CAPE, or the output of an existing ML model, was interpreted as the parameter having predictive power. Many ERA5 surface parameters were also examined. Ultimately, 40 variables were selected from an initial pool of over 60 candidate parameters (not listed). These parameters are defined in Table 1. Feature redundancy was a secondary consideration in the screening process, which means strong correlations between variables can be found (section 3.1), which is later considered in a feature selection experiment (section 3.2.2).

2.3. Classifiers

2.3.1. Logistic Regression

As a baseline multivariate model, we use logistic regression (James et al., 2013). Despite its simplicity, it often performs well on many real-world problems, including nonlinear problems in atmospheric science such as convection. For example, logistic regression outperformed random forests (RFs) in distinguishing between lightning and nonlightning days in Bates et al. (2018) and in predicting small-scale convective initiation from Numerical Weather Prediction (NWP) model and geostationary satellite data in Mecikalski et al. (2015).

2.3.2. Decision Trees

Two ML algorithm based on decision tree ensembles are included: RFs and boosted decision trees. These algorithms differ mainly on how the ensembles are generated. In a RF, decision trees are trained independently so that a random subsection of the data is used to train each tree. Furthermore, the trees are grown so that random subsections of features are considered at each node and the feature resulting in the best split is chosen. Often very deep trees are grown, resulting in individual trees suffering from overfitting and high variance. However, because each tree is fit to a different subsection of the data, the variance is reduced by averaging the predictions.

Boosting is an ensemble technique of training models sequentially, where each model aims to minimize the prediction errors of the previous model. The model is usually a "weak" learner, that is, a fairly simple model that does not perform well on its own. This framework of iteratively improving weak learners has been found to work particularly well for decision trees. After a suitable number of boosting iterations (which can be determined, e.g., by cross-validation), the final prediction is a weighted mean of all models. In this work, we use a LightGBM (LGBM), which is a computationally efficient Gradient Boosting Decision Tree algorithm (Ke et al., 2017).

RFs have been used with good results in many complex weather prediction applications, including predicting the initiation of mesoscale convective systems (Ahijevych et al., 2016), classifying storm types (Gagne et al., 2009), and recently to emulate a traditional convective parameterization in GCM simulations (Gorman & Dwyer, 2018). Gradient Boosting Decision Tree seems to have attracted little attention in atmospheric science but have been used successfully, for example, in solar energy prediction (McGovern et al., 2015). These models have proven highly effective for many real-world classification and regression problems, featuring in many winning solutions in data science competitions (Nielsen, 2016).

2.3.3. NN

For a technical description of NNs, the reader is referred to Chapter 5 in Bishop (2006). NNs are a class of ML algorithms which map inputs to outputs by one or more layer of *nodes* (also called neurons) connected to each other by nonlinear functions with adjustable parameters (weights). Thus, the input-output mapping represents a series of adjustable nonlinear transformations. During network training, the goal is to find a set of weights which minimize some measure of difference between the NN output and the training labels. In regression models, root-mean-square error can be used as the loss function. In a classification model, *cross-entropy error* should be preferred (equation 4.90 in Bishop, 2006). The loss function is then minimized using some variant of gradient descent, whereby a step, the magnitude of which is controlled by the learning rate, is taken in the direction of the steepest descent (the negative of the gradient) on each training iteration. The learning rate can be held constant or changed automatically after each epoch (a pass through all training data) by using an adaptive learning method. A commonly used optimizer is Adam, an adaptive stochastic gradient descent (SGD) algorithm with momentum (Kingma & Ba, 2014).

NNs can approximate any smooth nonlinear function (Hornik et al., 1989) but are difficult to interpret and computationally expensive to train. They can also be difficult to tune (e.g., finding a suitable complexity by adjusting the number of layers and neurons).

2.4. Metrics

The training data are characterized by a very low fraction of thunderstorm events (roughly 1% in Europe and 1.2% in Sri Lanka). Our problem is therefore a highly imbalanced two-class classification problem where positive samples, corresponding to detected lightning, are vastly outnumbered by negative samples. Class imbalance can be a serious issue when training classifiers, as a traditional performance metric such as accuracy could in this case be maximized by simply predicting all examples as the majority class (Liu et al., 2009a). Furthermore, misclassifying the minority class is often a more serious issue than misclassifying the majority class. For example, failing to issue a severe weather warning when one occurs (a false negative) is likely to have more serious consequences than issuing a false alarm (false positive).

Ultimately, if the costs corresponding to the four possible outcomes of yes/no forecasts are unknown, no scalar metric can be used to measure skill fully or optimally (see Wilks, 2006, section 7.2.3). A common way of evaluating binary classifiers is by plotting the receiving operating characteristic (ROC) curve. The ROC curve depicts the true positive rate against the false positive rate at different threshold values used to convert probabilistic model output into binary predictions. However, for class-imbalanced problems, the ROC curve can be misleading and a better option is to use the Precision-Recall (PR) curve, which depicts the true positive rate (recall) against precision (Saito & Rehmsmeier, 2015). We use the area under the PR-curve (PR-AU) to summarize model performance and optimize hyperparameters by maximizing this score, or the nearly equivalent average precision, with respect to a subset of the data withheld for model validation. The average precision of the validation data (valAP) is also used for early stopping, a generalization method whereby model training is stopped when the validation error begins to increase (a sign of the model overfitting to the training data).

2.5. ML Workflow

ML models for predicting TO in Europe and Sri Lanka were developed according to the procedure outlined below, aimed at tackling class imbalance and maximizing model performance.

1. The European data sets (May–August 2012–2017) were merged and the data divided into training (4/6), validation (1/6), and testing (1/6) subsets. Training data are used for model training, that is, fitting internal model parameters while validation data are used for optimizing model hyperparameters. The test data are used to obtain an unbiased estimate of performance. The data were divided in an interleaved manner: 2014 was reserved for testing, while the May, June, July, and August months were withheld from 2012, 2013, 2015, and 2017, respectively, for validation. The remaining data were used for training. Data from Sri Lanka (January–December 2016–2017) were similarly divided into training (2016), validation (even numbered months from 2017), and testing (odd numbered months from 2017) subsets.
2. To reduce class imbalance, the majority class (null cases with no lightning) in the training data was undersampled using an informed undersampling method available in Scikit-learn called BalanceCascade (Liu et al., 2009a). The goal was to undersample easily predictable null cases which dominate the data, for example, samples with zero CAPE. Essentially, a simple classifier is first trained (here, a Scikit-learn gradient-boosting classifier using LI_{MU}, *tciw*, $MRH_{600\text{–}800}$, and $CAPECIN_{\mathrm{MU}}$ as inputs) and correctly predicted majority samples are then removed according to a user-specified ratio of minority to majority samples. The following ratios were tested: 1–7, 1–9, and 1–20. Using a 1 to 9 ratio (10% thunderstorm frequency) led to the highest validation performance using LGBM (section 3.2). This reduced the training data set from 18.5 million to 1.9 million samples in Europe. Training data for Sri Lanka were also undersampled, but since almost all samples had positive CAPE, we chose to discard only roughly half of all null events (resulting in 1.3 million samples and 2.5% event probability).
3. RFs and logistic regression models were trained using all 40 input variables.
4. Boosted decision trees were trained using LGBM. First, Bayesian optimization was used to tune model hyperparameters. Recursive feature selection using the optimized model was then carried out in order to prune redundant features from the 40 initial features.
5. NNs were trained using the reduced feature set from the previous step and Bayesian optimization to tune the model architecture. Before training, all input data ware converted to a range between 0 and 1.
6. Data from Europe and Sri Lanka were pooled together, and NNs were trained to predict TO globally.
7. Model outputs were calibrated using Platt's scaling in order to correct for skewed class probabilities resulting from undersampling. This method is based on fitting a logistic regression model to the classifier outputs and true labels. The full training set was used to fit the model.
8. Models were evaluated using independent test data.

3. Model Training and Optimization

In ML, internal model parameters are learned directly from data during training. However, models also have various hyperparameters which control the complexity and regularization. These parameters need to be specified in advance and can strongly affect performance. In this section, we describe model training and hyperparameter optimization. Logistic regression and RFs were developed using the Scikit-learn ML library for Python (Pedregosa et al., 2011), while NNs were developed using Keras, a deep learning library for Python (https://keras.io/).

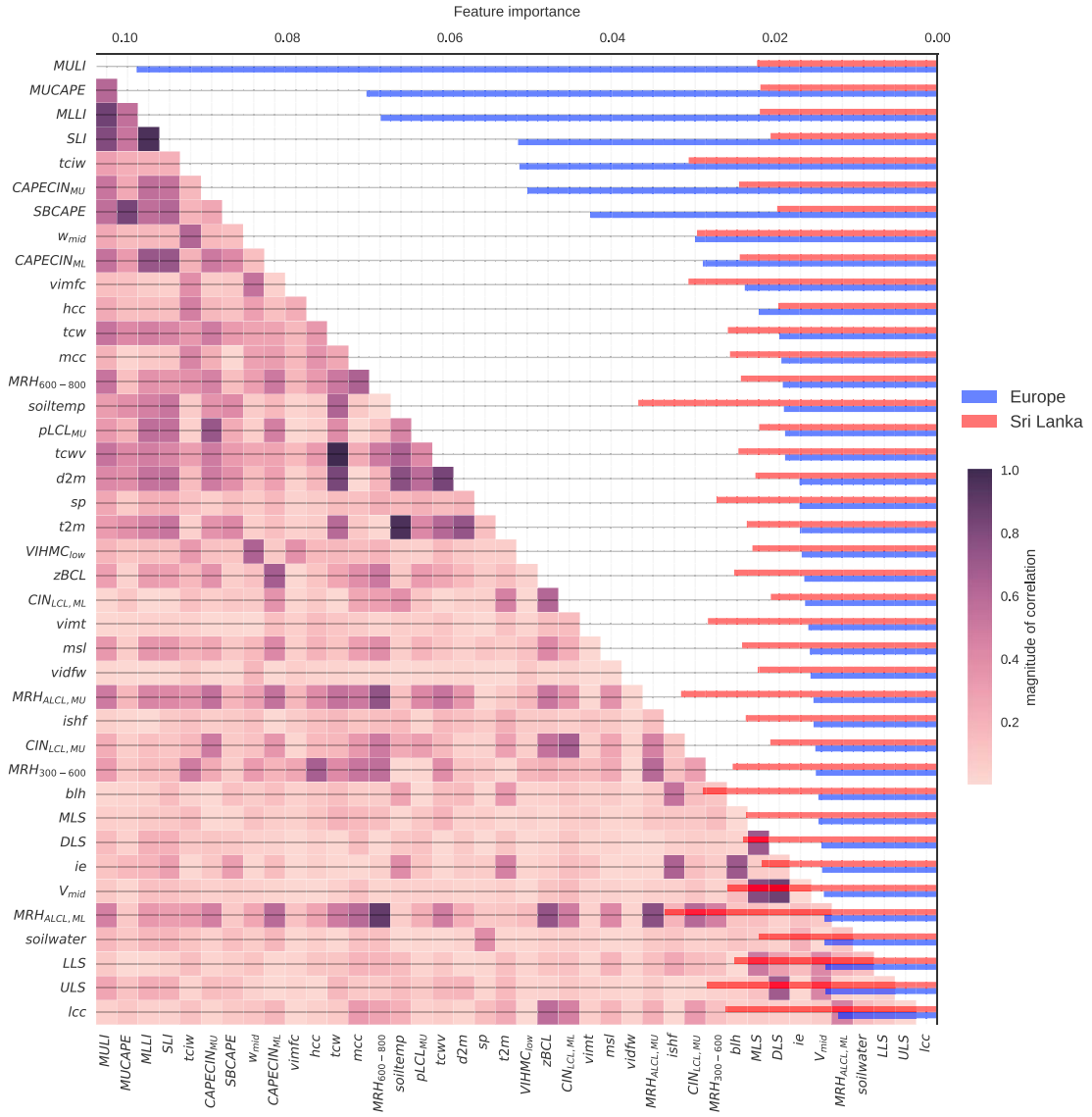

Figure 1. Correlation matrix depicting intercorrelation of features (lower left corner) and feature importance in a random forest model for Europe (blue bars) and Sri Lanka (red bars). The correlation matrix was computed using the training data from Europe.

3.1. Logistic Regression and RFs

Separate logistic regression models were fitted to the training data from Europe and Sri Lanka. Due to a large number of predictors, the Lasso regularization method (Ch. 6.2.2. in Osuri et al., 2017) was used with logistic regression to constrain the model to use an optimal subset of predictors. Five different values for the tuning parameter C, the inverse of regularization strength, were tested (0.01, 0.1, 1, 10, and 100). $C = 10$ led to the best performance in both domains with a valAP of 0.193 in Europe and 0.152 in Sri Lanka.

Next, we trained RFs, which are also relatively easy to tune. The most important hyperparameter is the maximum depth of trees. We used fully grown trees as restricting the tree depth degraded performance based on a quick test. Regarding the number of trees in the ensemble, it is considered good practice to simply grow as many trees as computationally feasible. An ensemble of 800 trees was found to be sufficient for obtaining good performance, with minimal gains for larger forests. All other hyperparameters were kept at their default value in Scikit-learn. RFs trained in this manner had valAP of 0.245 in Europe and 0.185 in Sri Lanka. A practical drawback of RF is a large memory requirement owing to the large tree depth and ensemble size; here the stored models took several gigabytes of disk space.

Table 2
The Hyperparameters of LGBM and Neural Networks Which Were Tuned Using Bayesian Optimization or by Hand for Selected NN Parameters, by Maximizing valAP

Hyperparameter	Explored space	Selected value
LGBM hyperparameters optimized using Bayesian optimization		
Maximum tree depth	3–25	13
Maximum number of leaves	15–100	81
Minimum data in leaf	20–120	64
Feature fraction	0.3–1	0.472
Minimum split gain	0.01–0.2	0.160
Minimum child weight	10–2,000	383.7
Neural network hyperparameters optimized using Tree Parzen Estimators		
Number of layers[a]	2, 3, 4, 5	3
Neurons in each layer[a]	30, 60, 90, 120, 150	120, 60, 30
Dropout rate after each layer[a]	0.0–0.4	0, 0.09, 0.15
Neural network hyperparameters tuned by hand		
Hidden neuron initialization	He, Glorot	Glorot
Hidden layer activation function	Exponential, Scaled Exponential, and Rectified Linear Unit (ELU, SELU, ReLU)	ELU
Optimizer (learning rate)	Adam (0.001), SGD with momentum (0.001, 0.01, 0.05, 0.1)	Adam
Class weight	1:1, 1:5, 1:10, 1:100	1:1
Batch size	256, 512, 1024	1024

Note. The explored range of parameters and the selected optimal values for Europe are given.
[a]Including the input layer and excluding the output layer. LGBM = LightGBM; NN = neural network.

A benefit of decision tree methods is that they are relatively easy to interpret by quantifying how much each feature in the model contributes to the overall predictive performance. In Scikit-learn, feature importance is computed as the associated (normalized) total reduction of the criteria used to measure the quality of a split, which in this case is the Gini impurity (Louppe et al., 2013). The feature importances for Europe and Sri Lanka are depicted as a bar chart in Figure 1. The results vary greatly with domain. In Europe, traditional instability measures such as CAPE and Lifted Index are associated with very high feature importance, with most of the predictive performance of the model brought by these few features. In Sri Lanka, these predictors are ineffective, and no single feature scores much higher than the rest. Interestingly, the most important feature for this domain is reported as soil temperature, which also came second in a boosted tree model. Whether this result is physical cannot be confirmed, but the significance of land surface conditions on the initiation of deep convection over the Indian Monsoon region has been noted previously in Osuri et al. (2017).

3.2. Boosted Trees

3.2.1. Hyperparameter Selection

The LGBM model has a large number of hyperparameters, the optimization of which using brute force is unfeasible. In order to find a good set of parameters in as few iterations as possible, we employ a Python implementation of Bayesian optimization (https://github.com/fmfn/BayesianOptimization) in which the objective function is modeled as a Gaussian process and the query points of the feature space are chosen in an informed manner (Snoek et al., 2012). The Bayesian optimization was initialized with 20 random points in the feature space and ran for 15 iterations. Only the European data were used for optimizing the model. The explored hyperparameter space and resulting optimal values are presented in Table 2. Note that the number of boosting iterations could be ignored by using early stopping (setting it at a high number). The optimized model had a maximum tree depth of 13 and a maximum of 81 leaves in each tree. Using the default learning rate (0.1), it took roughly 200 boosting iterations and only a few minutes of GPU running time to reach a maximum in valAP (0.271).

3.2.2. Feature Selection

The correlation matrix (Figure 1) revealed many strong pairwise correlations among features. Since removing redundant features will reduce unnecessary complexity and can occasionally improve performance,

we decided to use LGBM to examine the impact of removing features with low importance and/or strong correlation with other features.

Based on Figure 1, removing the following features was tested one by one: LI_{SFC}, $CAPE_{SFC}$, ULS, DLS, lcc, ie, $MRH_{ALCL,ML}$, $MRH_{300–600}$, $CIN_{LCL,MU}$, and $VIHMC_{low}$. In addition, we tested transforming two features: 2-m dew point temperature (d2m) into dew point depression and total column water (tcw) was subtracted with total column cloud ice water and total column water vapor, resulting in the sum of cloud liquid water, rain, and snow. On each feature trial 10 LGBM models were trained on the European data using different random initializations. A change was deemed beneficial and made permanent if it resulted in either the mean or the maximum valAP increasing or neither decreasing by more than 0.3%. This simple experiment, which did not explore different permutations of features, led to the removal of four redundant features (SLI, $CAPE_{SFC}$, $CIN_{LCL,MU}$, and DLS) and transforming d2m and tcw into new features. As a result, the maximum valAP increased from 0.271 to 0.275, with most of the improvement coming from feature transformations. Finally, an ensemble of 10 LGBM models with the same hyperparameters and inputs was trained on data from Sri Lanka. The best model had a validation score of 0.205 in this domain.

3.3. NNs

3.3.1. Hyperparameter Tuning

We use feedforward NNs for predicting thunderstorms, testing both shallow and deep NNs. To tune the NN architecture, we utilized the Python package Hyperopt (through the Keras wrapper Hyperas), which is based on a Bayesian optimization technique using Tree Parzen Estimators (Bergstra et al., 2011). The number of hidden layers, hidden neurons in each layer, and the strength of dropout regularization (Srivastava et al., 2014) were optimized using Hyperopt, while remaining hyperparameters were tuned manually. The explored hyperparameter space and selected values for Europe are given in Table 2. Architecture optimization was carried out separately for Sri Lanka, where the best model was also a deep NN with three hidden layers but with more hidden neurons in each layer. The valAP of the optimized model in Europe (Sri Lanka) was 0.277 (0.202) and therefore similar or slightly better compared to boosted decision trees.

3.3.2. A Global Model

As a final experiment, we pooled the training data from Europe and tropics together to see if a "global" model can be developed which performs better than models trained for specific domains. The data were aggregated so that the training and validation data sets used previously for each domain were simply merged. As a result, 41% of all training data but only 15% of positive samples are from Sri Lanka. The model architecture was optimized once again with Hyperopt and resulted in a more complex model with four hidden layers.

Once trained, the model was evaluated separately for each domain using the validation data from these regions. The global NN had a valAP of 0.270 in Europe and 0.204 in Sri Lanka. The performance was therefore slightly worse in Europe, but incrementally higher in Sri Lanka, compared to models trained previously. Given that tropical and middle-/high-latitude convection differ in many respects, the lack of improvement is not surprising. The input variables are furthermore likely to suffer from domain-dependent errors and biases which makes this a difficult goal. The similar level of performance for the global and specialized models should be considered at least a partial success; it is plausible that the new model would generalize well to climates in between the two regions.

4. Evaluation

4.1. Classifier Skill

To evaluate model skill for the purposes of forecasting, nowcasting, and triggering convection in large-scale models, we consider individual events of observed and predicted TO using hourly test data. Model skill is summarized in Figure 2 by using the areas under the PR and ROC curves, which are two complementary measures of skill obtained for a range of decision thresholds. In this evaluation we have included three traditional convective trigger schemes: regular (nondiluted) CAPE, dilute CAPE, and dilute dCAPE, which were previously evaluated alongside other schemes in Suhas and Zhang (2014). All three parameters were calculated using the most unstable parcel in the lowest 350 hPa and further mixed over a depth of 50 hPa for regular CAPE. Dilute CAPE accounts for dilution of the updraft by mixing entropy properties using a constant fractional entrainment rate (Neale et al., 2008). Finally, dCAPE refers to CAPE generation by large-scale advection. By using PR and ROC curves, the trigger parameters are not just evaluated at a single threshold, whose optimal values are in reality often domain dependent but are fixed in large-scale models

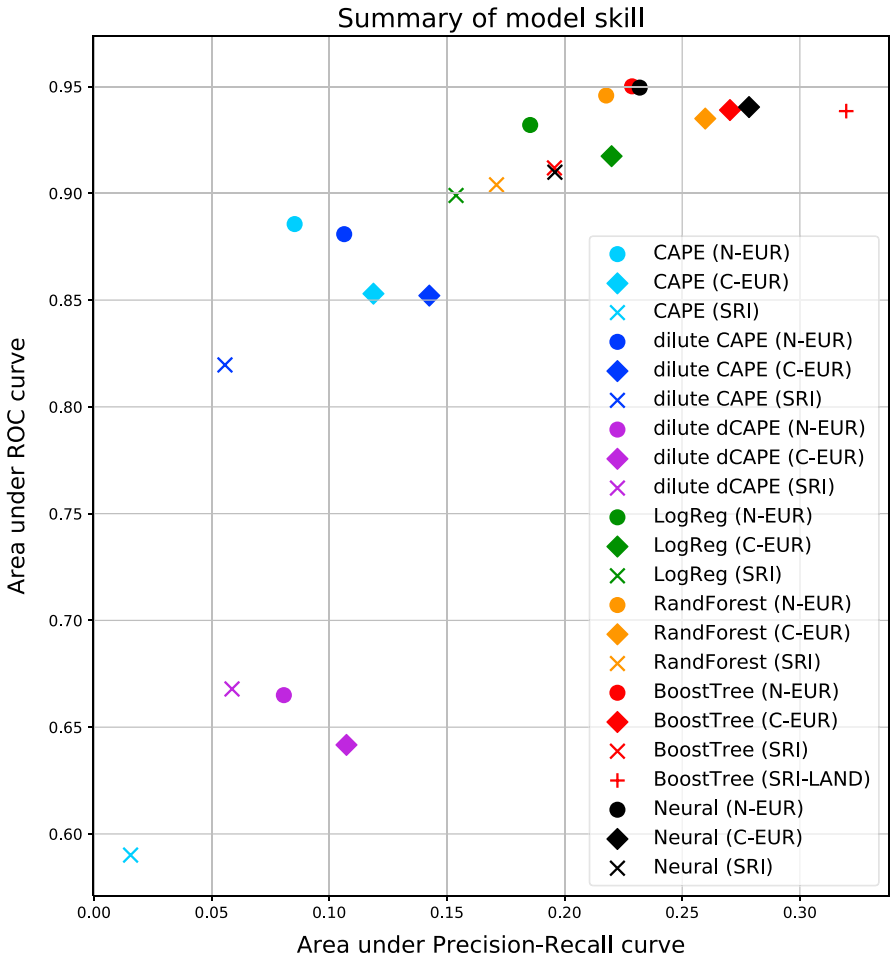

Figure 2. Area under the precision-recall (PR) and receiving operating characteristics (ROC) curves for different classifiers in Northern Europe (N-EUR), Central Europe (C-EUR), Sri Lanka (SRI), and a subdomain of the latter over land (SRI-LAND), obtained using independent test data.

(e.g., in the Community Atmosphere Model version 5, convection is triggered when dilute CAPE exceeds 70 J/kg).

Among the traditional parameters, regular CAPE performs relatively well in Europe but not Sri Lanka, and dilute dCAPE has poor performance everywhere. Our results seem incongruous with Suhas and Zhang (2014), who found the best performance for dilute dCAPE. However, the authors used different measures of skill, for example, the Equitable Threat Score. In our analysis dilute dCAPE had a higher maximum Equitable Threat Score than dilute CAPE in Sri Lanka (0.062 and 0.045, respectively; not shown) despite AU-ROC being much lower. Furthermore, we use lightning occurrence as the predictand, while previous evaluations have used precipitation.

The statistical models demonstrate much higher skill than any traditional parameter. Of these, NN and LGBM perform the best in all domains and have very similar skill. For conciseness we use only LGBM in further evaluation, choosing the simpler model of the two. In all figures presented hereafter, very similar results were obtained by using NNs instead of LGBM (not shown).

The low values of AU-PR relative to AU-ROC are explained by ROC being insensitive to the probability of false alarm (POFA), that is, the ratio of false alarms to all "yes" forecasts, which is in this instance high if a high recall (probability of detection) is also desired (Figure 3). The low precision (1-POFA) arises in this case from a very strict definition of what constitutes a thunderstorm event. For instance, in many cases a thunderstorm may be predicted but lightning is observed only after the 45-min window or in an adjacent grid cell, leading to a false alarm. In any case, the improvement over CAPE is substantial; for instance, if

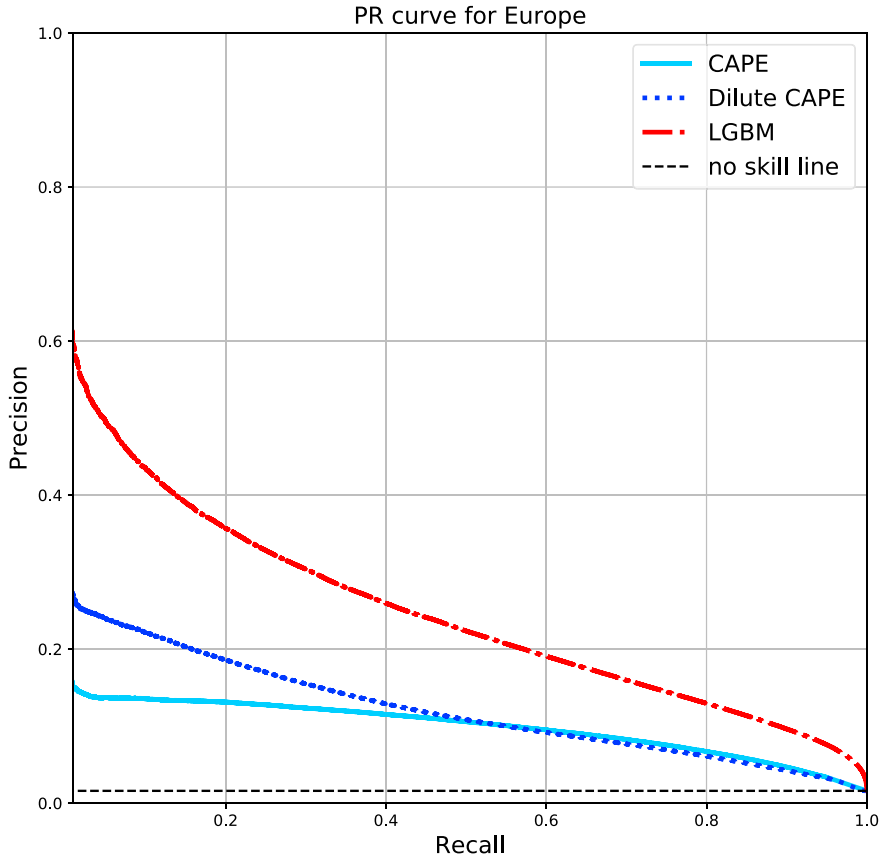

Figure 3. Precision-Recall curves for various classifiers using the test data from Europe. PR = Precision-Recall; CAPE = convective available potential energy; LGBM = LightGBM.

a probability of detection of 0.5 is desired, the probability of making a false alarm is more than doubled if CAPE is used instead of LGBM (Figure 3). The linear correlation coefficient between TO and model output likewise indicates a large improvement over traditional indices, reaching 0.37 for the LGBM output but only 0.22 for CAPE.

In Sri Lanka, the increase in skill brought by ML was even larger. While the scores were lower across the board compared to Europe, this was due to a large proportion of sea to land (Figure 5) and the skill being lower over sea. When considering a subset of the domain with mostly land, LGBM slightly outperformed NN and had a AU-PR (AU-ROC) of 0.320 (0.939). The best traditional parameter was again dilute CAPE with an AU-PR (AU-ROC) of 0.071 (0.803); although dilute dCAPE was better according to AU-PR (0.078), it received a much lower AU-ROC (0.652). Regular CAPE had no discernible relationship with observed TO in Sri Lanka with $r = 0.03$ and 0.06 for the full and land-only domains, respectively. Higher correlations were obtained using dilute CAPE (0.15 and 0.17) but still nowhere near the level obtained with ML (0.29 and 0.37 using LGBM).

Finally, the statistical significance of the results was assessed by a bootstrap experiment using the test data from N-EUR. A block bootstrap method was devised to account for the spatial and temporal correlation (available at https://github.com/peterukk/BlockBootstrap3D.jl). The bootstrapped distributions of $\mathrm{AUROC}_{\mathrm{LGBM}} - \mathrm{AUROC}_{\mathrm{diluteCAPE}}$ and $\mathrm{AUPR}_{\mathrm{LGBM}} - \mathrm{AUPR}_{\mathrm{diluteCAPE}}$ (using 8,000 bootstrap samples) were both more than 28σ above 0, meaning the null hypothesis of no skill difference can be rejected at any p value.

4.2. Diurnal Cycle

The successful application of ML in a convective parameterization depends not only on the accurate prediction of means but crucially also on important aspects of observed variability of convection being reproduced. Most convective parameterizations employed in large-scale models are based on an assumption of quasi-equilibrium between the large-scale forcing (which acts to generate CAPE) and the response

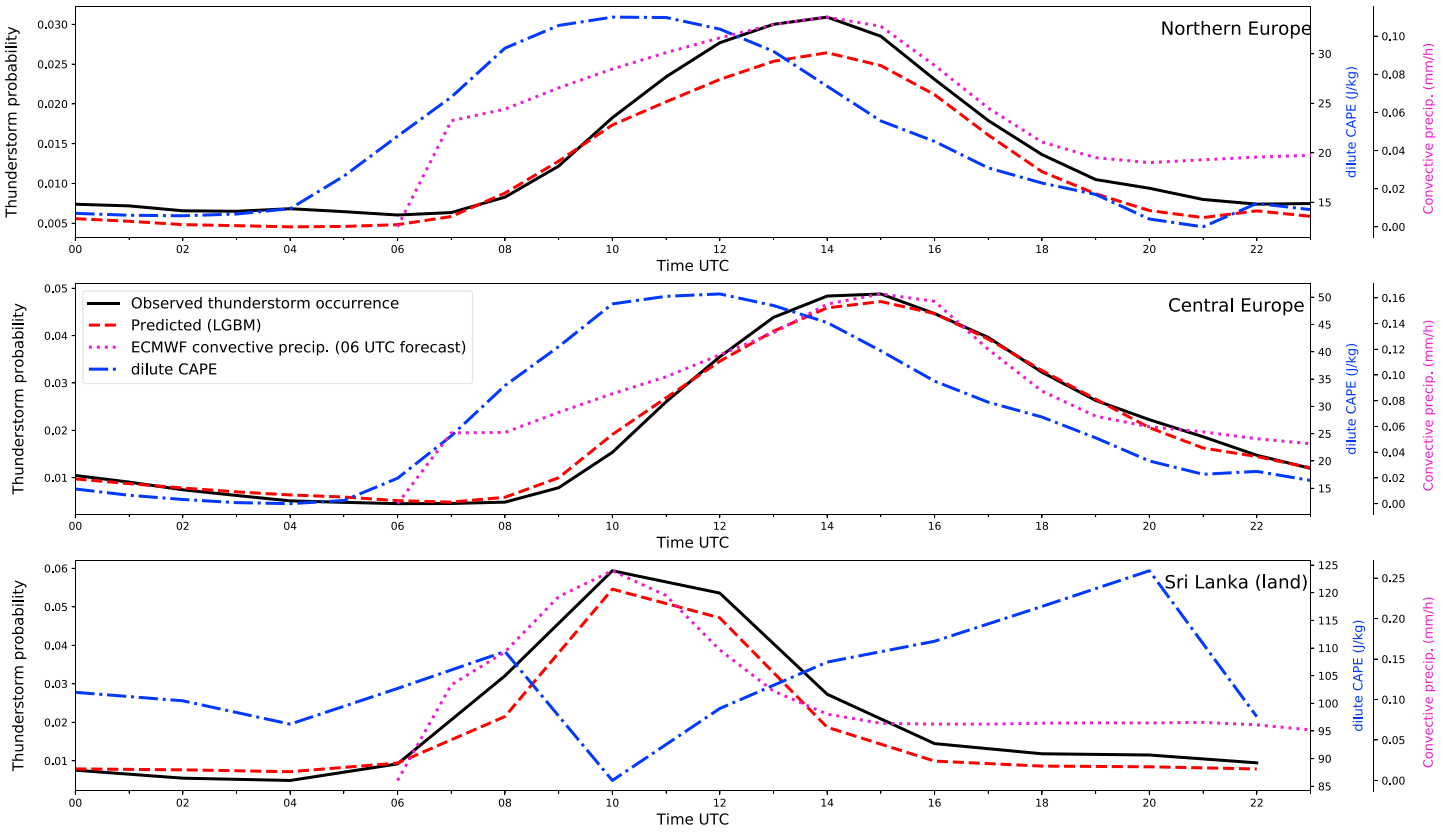

Figure 4. The diurnal cycle of observed thunderstorm occurrence, predicted thunderstorm occurrence by LGBM, and dilute CAPE as derived from the ERA5 hourly analysis, as well as 06 UTC forecast convective precipitation for hours 07–23 (by the ECMWF atmospheric model) for Northern Europe (top), Central Europe (middle), and the land-only domain for Sri Lanka (bottom), using the independent test data. All products are hourly except in Sri Lanka, where only the convective precipitation is hourly and the remaining variables were acquired at 2-hourly resolution. LGBM = LightGBM; CAPE = convective available potential energy.

to this forcing as realized convection (which consumes CAPE). While mass flux schemes which assume quasi-equilibrium can produce realistic middle-latitude synoptic variability and tropical wave spectra, the simulated diurnal cycle of convection over land is often several hours ahead of the observed cycle, a manifestation of nonequilibrium convection (Bechtold et al., 2014).

The diurnal cycle of observed and predicted deep convection (using TO as a proxy) as well as CAPE over different domains is displayed in Figure 4. Also shown is the 06-UTC forecast convective precipitation by the ECMWF model. ML predictions using LGBM match observations very well in all domains, although in N-EUR the peak is slightly underestimated. While the phase of the diurnal cycle of CAPE is many hours ahead of observed deep convection as expected, the ECMWF convective parameterization performs very well in terms of the phase of precipitation matching that of observed lightning. This is due to an update in 2013 to the closure used in the IFS convective parameterization which made it possible to represent nonequilibrium convection realistically (Bechtold et al., 2014). While Figure 4 suggests that the deep convective activity in the morning is still somewhat overestimated by the ECMWF model, this may be caused by comparing lightning occurrence to the intensity of convective precipitation, which are not the same thing. Nevertheless, the same can be seen also in an evaluation of the scheme over Europe and the Sahel region in Figure 5 in Bechtold et al. (2014). This suggests that there may be room for improvement in the ECMWF parameterization. While the results are certainly promising, it should be stressed that they are not based on a prognostic validation. A reanalysis-based validation may in principle skew the results in favor of multivariate methods, for example, when satellite observations of long-lived convective systems are assimilated. However, we found that the relative improvement in AU-PR between dilute CAPE and LGBM was only slightly decreased when cases with previous lightning activity were removed.

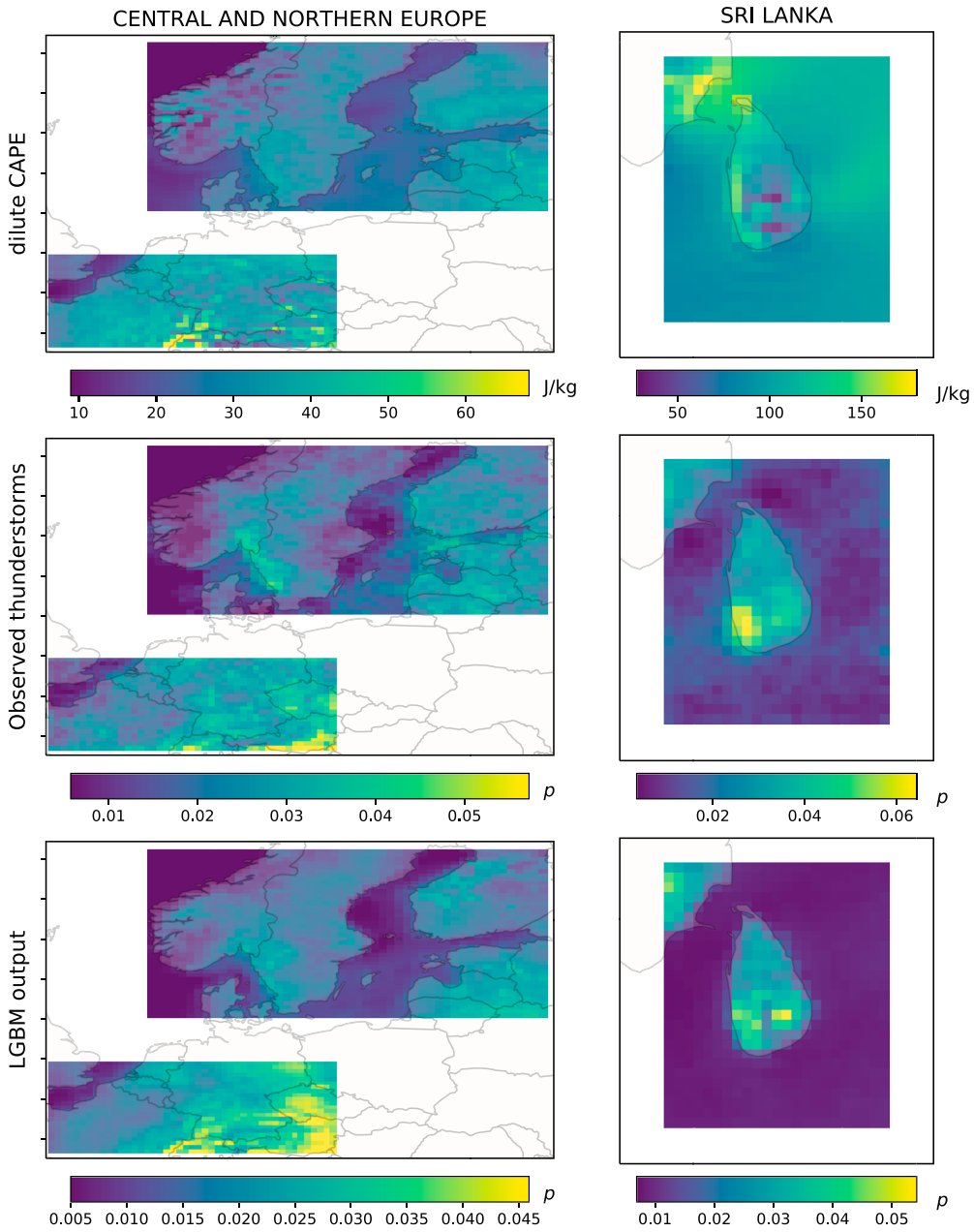

Figure 5. The grid box means of dilute CAPE (top row), observed hourly thunderstorm occurrence (middle row) and predicted 45-min thunderstorm occurrence by LGBM (bottom row) in Europe (left) and Sri Lanka (right) using test data in Europe (May–August 2014, hourly data) and the test and validation data in Sri Lanka (2017, 2-hourly data). CAPE = convective available potential energy; LGBM = LightGBM.

4.3. Correlation With Observed Thunderstorms on Larger Scales

Grid box means of observed TO and ERA5-derived convective predictors in Europe (2014, test data) and in Sri Lanka (2017, validation and test data) are plotted in Figure 5. Regions of high thunderstorm activity seem to correspond well with ML predictions in Europe. CAPE is elevated over the Baltic Sea relative to the observed and predicted thunderstorm activity. In Central Europe, ML predictions are clearly better in the east and southeast, where CAPE underpredicts convection. In Sri Lanka, the evaluation is hampered by a small domain, but the LGBM output matches observed deep convection better than CAPE or the forecast convective precipitation (not shown) which were both elevated over sea.

Table 3
Linear Correlation Between Area Means of Monthly Means of Parameters in Europe (May–August 2012–2017) and Sri Lanka (January–December 2016–2017), Where TO Refers to the Observed (45 min) Thunderstorm Occurrence, vimfc Is the Vertical Integral of Divergence of Moisture Flux, y_{LGBM} Is the Output of the Boosted Decision Tree Model, cp_{fc} Is the Forecast Accumulated Convective Precipitation and C-G Flashes Is the Number of Observed Cloud-to-Ground Flashes

y_1, y_2	N-EUR$_{sub}$	C-EUR	Sri Lanka
TO, CAPE	0.967	0.802	0.374
TO, dilute CAPE	0.905	0.851	0.417
TO, total column ice water	−0.332	0.074	0.670
TO, vimfc	0.155	−0.173	−0.728
TO, cp_{fc}	0.628	0.694	0.725
TO, y_{LGBM}	**0.989**	**0.951**	**0.944**
C-G flashes, CAPE	0.9361	0.772	—
C-G flashes, cp_{fc}	0.560	0.491	—
C-G flashes, y_{LGBM}	**0.959**	0.783	—
C-G flashes, $y_{LGBM}\times$ CAPE	0.949	**0.858**	—

Note. All means have been calculated using synoptic 2-hourly analyses, except cp_{fc} which was obtained as a daily mean created from contiguous hourly forecast data (initialized twice a day). CAPE refers to CAPE$_{MU}$, except for C-EUR, where CAPE$_{SFC}$ was used due to lower skill of CAPE$_{MU}$ in the Alpine region. Only correlations between the European domains are comparable due to similar surface areas, obtained by using a northern subsection of N-EUR (N-EUR$_{sub}$). The highest correlations for each domain and predictand are bolded.

Next, the regional means of monthly means of parameters were compared to observed TO and CG flash totals in different regions. A strong relationship would enable the use of reanalysis data to infer climate trends in the occurrence convective storms over the past 40 years or more. Linear correlation coefficients between monthly values were calculated using all available data, which amounts to 24 months in each domain (Table 3). In Northern Europe, the correlation between monthly mean CAPE and observed thunderstorm frequency is already remarkably high at 0.967, with limited room for improvement. With LGBM, the correlation reaches 0.989. In C-EUR, CAPE has a lower correlation with observed monthly thunderstorm activity (r = 0.80–0.85). ML predictions correlate strongly with observations in both C-EUR (0.95) and Sri Lanka (0.94). The high correlations suggest that ERA5 and ML could be used to reconstruct past thunderstorm activity globally. Individual predictors derived from ERA5 were less successful in Sri Lanka, where CAPE correlates poorly with storm frequency also on larger scales (r = 0.37–0.42). Moisture flux convergence may be a better predictor for storm initiation in the tropics (r = 0.73).

Finally, monthly total number of C-G flashes in Northern Europe has a strong correlation with CAPE (0.94) and ML-predicted storm occurrence (0.96). In C-EUR, the best predictor for monthly C-G flashes is the product of CAPE and ML model output. A likely explanation is that CAPE gives a theoretical upper bound for the intensity of deep convection (for which C-G flash frequency is a proxy) but is conditional to convective initiation, the probability of which is modeled by LGBM. The data used for Sri Lanka did not differentiate between C-G and intracloud flashes, but the correlation between CAPE and mean monthly flash density was 0.65–0.69 (not shown).

5. Environmental Factors Regulating Deep Convection

The use of ML, lightning, and reanalysis data offers an opportunity to assess the relative importance of different variables for deep moist convection in different climates. Caveats here include the correlation of features affecting the feature ranking and uncertainties associated with the use of reanalysis data. Specifically, the results may be sensitive to variable- and domain-dependent errors in the reanalysis. Nevertheless, the use of a nonlinear statistical method and observations of lightning (which unlike precipitation, is always associated with deep, moist convection) may be valuable for gaining insight into the environmental factors controlling convection in the tropics and midlatitudes. Identifying such control variables observationally

has been difficult according to Yano et al. (2013), who review this subject in the context of convective parameterization.

First, in order to gain a more reliable estimate of feature importance, we conducted a further experiment using the ELI5 python library (http://eli5.readthedocs.io) to measure the *permutation feature importance* of the final NN and LGBM models. The method is based on measuring the decrease in score (here, valAP) when the values of a feature are shuffled. For each domain, we then took the mean of the feature importances for the two models (not shown). Overall, the ranking was similar to Figure 1, with stability indices having high importance in Europe but not in the tropical region. One difference was that soil temperature received a high score also in Europe, coming third after LI_{MU} and LI_{ML}. In Sri Lanka, the three most important features were soil temperature, soil water content, and the sum of cloud liquid water, rain, and snow. Low-level convergence and w_{mid} had a low score in both domains.

From our results we draw the following key findings:

- In Sri Lanka, regular (undilute) CAPE has a weak correlation with observed TO. This is consistent with studies showing that tropical convection does not have any clear correlation with CAPE (e.g., Sherwood, 1999).
- Yano et al. (2013) stress that convection in the tropics, especially over sea, is instead controlled by humidity in the low-to-middle troposphere. The larger relative importance of humidity arises due to weak horizontal temperature gradients in the tropics which lead to low variability of undiluted CAPE. Our results are in agreement with this: in Sri Lanka humidity variables had high feature importances, and accounting for updraft dilution was very important.
- Column-integrated moisture and humidity in the low-to-middle troposphere were important thunderstorm predictors also in Europe. For example, relative humidity in the 600- to 800-hPa layer greatly increased the probability of lightning, also for large values of CAPE. This parameter may be a simple yet effective way of accounting for the effect of environmental humidity on deep convection through mixing and the associated shallow-to-deep convection transition (Ukkonen et al., 2017; Wu et al., 2009).
- The importance of CIN for observed deep convection (or lack thereof; see section 5.5 Yano et al., 2013) is difficult to establish. We did not include CIN as it is traditionally defined below the LFC. CIN below the LCL, and the sum of CAPE and CIN in a 250-hPa-thick layer above the LCL, performed fairly well in both domains. In Sri Lanka, thunderstorm probability increased monotonically with the latter parameter but not with undilute CAPE.
- According to Mapes (1997), the importance of CIN and other "low-level controls" on deep convection depends on the scale of the convective systems, so that convection on larger scales responds to changes in instability by large-scale processes and CIN is unimportant. On the mesoscales, the spatial organization of convection (into arcs or lines) is governed by PBL processes which act to reduce CIN and provide low-level lift by which CIN can be overcome. From this, it should follow that the monthly mean convective activity over large areas has a high correlation with CAPE, but individual thunderstorms can be more skillfully predicted by a multivariate model. This was clearly the case in Northern Europe, where CAPE was sufficient to capture monthly thunderstorm frequency. However, In Central Europe and especially Sri Lanka, ML predictions had a much higher correlation with monthly TO than any large-scale variable. This suggests that estimates of future changes in convective storms need to also consider storm initiation (governed by many factors) and not only changes in conditional instability.
- Thunderstorm probability for the same very favorable values of CAPE, CIN, and $MRH_{600–800}$ was much higher in Europe than in Sri Lanka. Assuming that this result is physical, a possible explanation may be the temperature dependence of buoyancy reduction through mixing. For similar values of relative humidity, the magnitude of evaporation occurring when updraft air is mixed with the environment is much larger in higher temperatures.

6. Discussion and Conclusion

In this paper, we have explored the use of ML to predict the occurrence of deep convection in different climates. Models were trained using lightning data and a high-resolution global reanalysis. Considerable effort was devoted to improve model performance, generalization, and interpretation. First, atmospheric profiles were not used directly as inputs; instead, a large number of convective parameters were calcu-

lated and evaluated. Second, Bayesian optimization was utilized for efficient and systematic tuning of hyperparameters.

Classifier skill for short-term thunderstorm predictions (0–45 min), as measured by the area under the PR-curve, was more than doubled in Europe by using NNs or boosted trees instead of CAPE. These models performed very well also in Sri Lanka, where convective indices did not. The relationship between undilute CAPE and TO was very weak in this region. Regarding the choice of model, we note that decision tree ensembles were able to offer similar performance to deep NNs, probably due to successful feature engineering.

Our results suggest several promising future applications. First, findings regarding the most skillful algorithms and predictors for short-term prediction of convective initiation should be relevant for nowcasting purposes. Second, the realistic diurnal cycle of ML-predicted TO in both Europe and Sri Lanka indicates feasibility for predicting the onset of deep convection in the context of convective parameterization. An ML-based convective trigger could rectify biases in the diurnal cycle of precipitation, which have been found in CMIP5 models (Harding et al., 2013) as well as in a new-generation Earth System Model (Zhao et al., 2018). Since ML classifiers predict class probabilities, one could be implemented as a stochastic triggering parameterization in a GCM, potentially improving the variability of simulated convection in tropical regions (Rochetin et al., 2014).

Finally, an almost perfect linear relationship was found between area-mean monthly TO and predictions using ML and ERA5 data. In Northern Europe, CAPE alone was able to explain the monthly variability. The results suggest that reanalyses and ML could be used to study climate trends in convective weather around the globe. This is important, as the impact of climate change on storm initiation is an unanswered question (Allen, 2018). While the sample used was small (24 months), the high performance of a single NN across different climates (section 3.2.2) suggests applicability also for longer time periods. Despite NN being unable to extrapolate to completely new climates, there is evidence that they can interpolate in between extremes (Rasp et al., 2018). Similarly, RFs in Gorman and Dwyer (2018) were able to generalize to a new regional climate as long as similar temperatures from a different region had been sampled during training. A potentially bigger issue with reconstructing past convective weather in this manner may be the impact of changes to the observing system, which inevitably introduce some uncertainty to calculated trends. However, such issues are alleviated by variational bias correction and other significant efforts to minimize nonclimatic influences in reanalyses; indeed, the primary aim of reanalysis has always been to provide a homogeneous record of the atmosphere (Dee et al., 2011). In any case, this issue is much more serious for direct records of convective weather.

Recent studies show that ML models are able to learn the underlying physical relationships governing subgrid convection effectively and are computationally efficient once trained. While challenges remain, subgrid parameterizations learned from CRM or convection-permitting RCM data could greatly improve GCMs in the coming years. In the future, parameterizations could be learned directly and sequentially (online learning) from assimilated observations and replace or complement traditional schemes. While advances in learning algorithms and data assimilation may in some cases be needed before the key physical variables in parameterizations can be assimilated and learned, it would already be possible to implement a classification model like ours which learns to predict lightning occurrence in a global NWP model.

Acknowledgments
We thank Marja Bister for helpful discussions on deep convection. The data and Python code for machine learning training and evaluation are available at zenodo.org (Ukkonen & Mäkelä, 2018). The Julia package for calculating convective indices from sounding data can be found online (https://github.com/peterukk/ConvectiveIndices.jl). This work was funded by the State Nuclear Waste Management Fund in Finland through the EXWE project (Extreme weather and nuclear power plants) of the Finnish Nuclear Power Plant Safety Research Programme 2015–2018 (SAFIR2018) and by Business Finland and the Ministry of Foreign Affairs of Finland through the Severe Storm Warning Service for Sri Lanka (SSWSS) project. Finally, we thank EUCLID for the European lightning location data used in this study.

References

Ahijevych, D., Pinto, J. O., Williams, J. K., & Steiner, M. (2016). Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Weather and Forecasting*, *31*(2), 581–599. https://doi.org/10.1175/waf-d-15-0113.1

Allen, J. T. (2018). Climate change and severe thunderstorms: Oxford Research Encyclopedia of Climate Science. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228620.013.62

Ávila, E. E., Bürgesser, R. E., Castellano, N. E., Collier, A. B., Compagnucci, R. H., & Hughes, A. R. W. (2010). Correlations between deep convection and lightning activity on a global scale. *Journal of Atmospheric and Solar-Terrestrial Physics*, *72*(14-15), 1114–1121. https://doi.org/10.1016/j.jastp.2010.07.019

Ban, N., Schmidli, J., & Schär, C. (2014). Evaluation of the convection-resolving regional climate modeling approach in decade-long simulations. *Journal of Geophysical Research: Atmospheres*, *119*, 7889–7907. https://doi.org/10.1002/2014jd021478

Bates, B. C., Dowdy, A. J., & Chandler, R. E. (2018). Lightning prediction for Australia using multivariate analyses of large-scale atmospheric variables. *Journal of Applied Meteorology and Climatology*, *57*(3), 525–755. https://doi.org/10.1175/jamc-d-17-0214.1

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J.-P., Beljaars, A., & Bormann, N. (2014). Representing equilibrium and nonequilibrium convection in large-scale models. *Journal of the Atmospheric Sciences*, *71*(2), 734–753. https://doi.org/10.1175/jas-d-13-0163.1

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pp. 2546–2554.

Bishop, C. M. (2006). *Pattern recognition and machine learning (information science and statistics)*. Berlin, Germany: Springer-Verlag.

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*, 6289–6298. https://doi.org/10.1029/2018gl078510

Brooks, H. E. (2013). Severe thunderstorms and climate change. *Atmospheric Research*, *123*, 129–138. https://doi.org/10.1016/j.atmosres.2012.04.002

Cesana, G., & Waliser, D. E. (2016). Characterizing and understanding systematic biases in the vertical structure of clouds in CMIP5/CFMIP2 models. *Geophysical Research Letters*, *43*, 10,538–10,546. https://doi.org/10.1002/2016gl070515

Dee, D. P., Källén, E., Simmons, A. J., & Haimberger, L. (2011). Comments on 'reanalyses suitable for characterizing long-term trends'. *Bulletin of the American Meteorological Society*, *92*(1), 65–70. https://doi.org/10.1175/2010bams3070.1

Dirmeyer, P. A., Cash, B. A., Kinter, J. L., Jung, T., Marx, L., Satoh, M., et al. (2011). Simulating the diurnal cycle of rainfall in global climate models: Resolution versus parameterization. *Climate Dynamics*, *39*(1-2), 399–418. https://doi.org/10.1007/s00382-011-1127-9

Dotzek, N., Groenemeijer, P., Feuerstein, B., & Holzer, A. M. (2009). Overview of ESSL's severe convective storms research using the European Severe Weather Database ESWD. *Atmospheric Research*, *93*(1-3), 575–586. https://doi.org/10.1016/j.atmosres.2008.10.020

Gagne, D. J., McGovern, A., & Brotzge, J. (2009). Classification of convective areas using decision trees. *Journal of Atmospheric and Oceanic Technology*, *26*(7), 1341–1353. https://doi.org/10.1175/2008jtecha1205.1

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*, 5742–5751. https://doi.org/10.1029/2018gl078202

Gorman, P. A. O, & Dwyer, J. G. (2018). 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

Harding, K. J., Snyder, P. K., & Liess, S. (2013). Use of dynamical downscaling to improve the simulation of Central U.S. warm season precipitation in CMIP5 models. *Journal of Geophysical Research: Atmospheres*, *118*, 12,522–12,536. https://doi.org/10.1002/2013jd019994

Herman, M. J., & Kuang, Z. (2013). Linear response functions of two convective parameterization schemes. *Journal of Advances in Modeling Earth Systems*, *5*, 510–541. https://doi.org/10.1002/jame.20037

Hersbach, H., & Dee, D. (2016). ERA5 reanalysis is in production. *ECMWF newsletter*, *147*(7).

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Houston, A. L., & Niyogi, D. (2007). The sensitivity of convective initiation to the lapse rate of the active cloud-bearing layer. *Monthly Weather Review*, *135*(9), 3013–3032. https://doi.org/10.1175/mwr3449.1

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York, United States: Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pp. 3146–3154.

Kendon, E. J., Ban, N., Roberts, N. M., Fowler, H. J., Roberts, M. J., Chan, S. C., et al. (2017). Do convection-permitting regional climate models improve projections of future precipitation change? *Bulletin of the American Meteorological Society*, *98*(1), 79–93. https://doi.org/10.1175/bams-d-15-0004.1

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Knist, S., Goergen, K., & Simmer, C. (2018). Evaluation and projected changes of precipitation statistics in convection-permitting WRF climate simulations over Central Europe. *Climate Dynamics*, 1–17. https://doi.org/10.1007/s00382-018-4147-x

Kyselý, J., Rulfová, Z., Farda, A., & Hanel, M. (2015). Convective and stratiform precipitation characteristics in an ensemble of regional climate model simulations. *Climate Dynamics*, *46*(1-2), 227–243. https://doi.org/10.1007/s00382-015-2580-7

Lin, J.-L., Lee, M.-I., Kim, D., Kang, I.-S., & Frierson, DarganM. W. (2008). The impacts of convective parameterization and moisture triggering on AGCM-simulated convectively coupled equatorial waves. *Journal of Climate*, *21*(5), 883–909. https://doi.org/10.1175/2007jcli1790.1

Liu, Y., Guo, L., Wu, G., & Wang, Z. (2009b). Sensitivity of ITCZ configuration to cumulus convective parameterizations on an aqua planet. *Climate Dynamics*, *34*(2-3), 223–240. https://doi.org/10.1007/s00382-009-0652-2

Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2009a). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539–550. https://doi.org/10.1109/tsmcb.2008.2007853

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems*, pp. 431–439.

MacGorman, D. R., Apostolakopoulos, I. R., Lund, N. R., Demetriades, N. W. S., Murphy, M. J., & Krehbiel, P. R. (2011). The timing of cloud-to-ground lightning relative to total lightning activity. *Monthly Weather Review*, *139*(12), 3871–3886. https://doi.org/10.1175/mwr-d-11-00047.1

Mapes, B. E. (1997). *Equilibrium vs. activation control of large-scale variations of tropical deep convection. In the physics and parameterization of moist atmospheric convection (pp. 321–358)*. Netherlands: Springer. https://doi.org/10.1007/978-94-015-8828-7_13

McGovern, A., Gagne, D. J., Basara, J., Hamill, T. M., & Margolin, D. (2015). Solar energy prediction: An international contest to initiate interdisciplinary research on compelling meteorological problems. *Bulletin of the American Meteorological Society*, *96*(8), 1388–1395. https://doi.org/10.1175/bams-d-14-00006.1

Mecikalski, J. R., Williams, J. K., Jewett, C. P., Ahijevych, D., LeRoy, A., & Walker, J. R. (2015). Probabilistic 01-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *Journal of Applied Meteorology and Climatology*, *54*(5), 1039–1059. https://doi.org/10.1175/jamc-d-14-0129.1

Mäkelä, A., Enno, S.-E., & Haapalainen, J. (2014b). Nordic lightning information system: Thunderstorm climate of Northern Europe for the period 2002–2011. *Atmospheric Research*, *139*, 46–61. https://doi.org/10.1016/j.atmosres.2014.01.008

Mäkelä, A., Shrestha, R., & Karki, R. (2014a). Thunderstorm characteristics in Nepal during the pre-monsoon season 2012. *Atmospheric Research*, *137*, 91–99. https://doi.org/10.1016/j.atmosres.2013.09.012

Neale, R. B., Richter, J. H., & Jochum, M. (2008). The impact of convection on ENSO: From a delayed oscillator to a series of events. *Journal of Climate*, *21*(22), 5904–5924. https://doi.org/10.1175/2008jcli2244.1

Nielsen, D. (2016). Tree boosting with XGBoost–Why does XGBoost win every machine learning competition? (Master's Thesis). Norwegian University of Science and Technology. https://brage.bibsys.no/xmlui/handle/11250/2433761

Osuri, K. K., Nadimpalli, R., Mohanty, U. C., Chen, F., Rajeevan, M., & Niyogi, D. (2017). Improved prediction of severe thunderstorms over the Indian Monsoon region using high-resolution soil moisture and temperature initialization. *Scientific Reports*, *7*(1). https://doi.org/10.1038/srep41377

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pohjola, H., & Mäkelä, A. (2013). The comparison of GLD360 and EUCLID lightning location systems in Europe. *Atmospheric Research*, *123*, 117–128. https://doi.org/10.1016/j.atmosres.2012.10.019

Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., et al. (2015). A review on regional convection-permitting climate modeling: Demonstrations prospects, and challenges. *Reviews of Geophysics*, *53*, 323–361. https://doi.org/10.1002/2014rg000475

Púčik, T., Groenemeijer, P., Rädler, A. T., Tijssen, L., Nikulin, G., Prein, A. F., et al. (2017). Future changes in European severe convection environments in a regional climate model ensemble. *Journal of Climate*, *30*(17), 6771–6794. https://doi.org/10.1175/jcli-d-16-0777.1

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Rochetin, N., Grandpeix, J.-Y., Rio, C., & Couvreux, F. (2014). Deep convection triggering by boundary layer thermals. Part II: Stochastic triggering parameterization for the LMDZ GCM. *Journal of the Atmospheric Sciences*, *71*(2), 515–538. https://doi.org/10.1175/jas-d-12-0337.1

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Schulz, W. (2005). Cloud-to-ground lightning in Austria: A 10-year study using data from a lightning location system. *Journal of Geophysical Research*, *110*, D09101. https://doi.org/10.1029/2004jd005332

Sherwood, S. C. (1999). Convective precursors and predictability in the Tropical Western Pacific. *Monthly Weather Review*, *127*(12), 2977–2991. https://doi.org/10.1175/1520-0493(1999)127<2977:cpapiti>2.0.co;2

Sherwood, S. C., Bony, S., & Dufresne, J.-L. (2014). Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, *505*(7481), 37–42. https://doi.org/10.1175/1520-0493(1999)127<2977:cpapiti>2.0.co;2

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pp. 2951–2959.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Suhas, E., & Zhang, G. J. (2014). Evaluation of trigger functions for convective parameterization schemes using observations. *Journal of Climate*, *27*(20), 7647–7666. https://doi.org/10.1175/jcli-d-13-00718.1

Taszarek, M., Brooks, H. E., & Czernecki, B. (2017). Sounding-derived parameters associated with convective hazards in Europe. *Monthly Weather Review*, *145*(4), 1511–1528. https://doi.org/10.1175/mwr-d-16-0384.1

Taszarek, M., Brooks, H. E., Czernecki, B., Szuster, P., & Fortuniak, K. (2018). Climatological aspects of convective parameters over Europe: A comparison of ERA-Interim and sounding data. *Journal of Climate*, *31*(11), 4281–4308. https://doi.org/10.1175/jcli-d-17-0596.1

Tawfik, A. B., & Dirmeyer, P. A. (2014). A process-based framework for quantifying the atmospheric preconditioning of surface-triggered convection. *Geophysical Research Letters*, *41*, 173–178. https://doi.org/10.1002/2013gl057984

Tawfik, A. B., Lawrence, D. M., & Dirmeyer, P. A. (2017). Representing subgrid convective initiation in the Community Earth System Model. *Journal of Advances in Modeling Earth Systems*, *9*, 1740–1758. https://doi.org/10.1002/2016ms000866

Ukkonen, P., Manzato, A., & Mäkelä, A. (2017). Evaluation of thunderstorm predictors for Finland using reanalyses and neural networks. *Journal of Applied Meteorology and Climatology*, *56*(8), 2335–2352. https://doi.org/10.1175/jamc-d-16-0361.1

Ukkonen, P., & Mäkelä, A. (2018). Data and code for training and evaluating machine learning models for thunderstorm prediction from reanalysis data, https://doi.org/10.5281/zenodo.2148541

Westermayer, A. T., Groenemeijer, P., Pistotnik, G., Sausen, R., & Faust, E. (2017). Identification of favorable environments for thunderstorms in reanalysis data. *Meteorologische Zeitschrift*, *26*(1), 59–70. https://doi.org/10.1127/metz/2016/0754

Wilks, D. (2006). Statistical methods in the atmospheric sciences. Elsevier Science, https://books.google.fi/books?id=_vSwyt8_OGEC

Williams, E. R. (2001). The electrification of severe storms. In C.A. Doswell (Ed.), *Severe convective storms* (Vol. 28, pp. 527–561). Boston, MA: American Meteorological Society. https://doi.org/10.1007/978-1-935704-06-513

Wu, C.-M., Stevens, B., & Arakawa, A. (2009). What controls the transition from shallow to deep convection? *Journal of the Atmospheric Sciences*, *66*(6), 1793–1806. https://doi.org/10.1175/2008jas2945.1

Xie, S. (2004). Impact of a revised convective triggering mechanism on Community Atmosphere Model Version 2, simulations: Results from short-range weather forecasts. *Journal of Geophysical Research*, *109*, D14102. https://doi.org/10.1029/2004JD004692

Yano, J.-I., Bister, M., Fuchs, Z., Gerard, L., Phillips, V. T. J., Barkidija, S., & Piriou, J.-M. (2013). Phenomenology of convection-parameterization closure. *Atmospheric Chemistry and Physics*, *13*(8), 4111–4131. https://doi.org/10.5194/acp-13-4111-2013

Yano, J.-I., Liu, C., & Moncrieff, M. W. (2012). Self-organized criticality and homeostasis in atmospheric convective organization. *Journal of the Atmospheric Sciences*, *69*(12), 3449–3462. https://doi.org/10.1175/jas-d-12-069.1

Yano, J.-I., Soares, P. M. M., Köhler, M., & Deluca, A. (2015). The convective parameterization problem: Breadth and depth. *Bulletin of the American Meteorological Society*, *96*(8), ES127–ES130. https://doi.org/10.1175/bams-d-14-00134.1

Zhao, M., Golaz, J.-C., Held, I., Guo, H., Balaji, V., Benson, R., et al. (2018). The GFDL global atmosphere and land model AM4. 0/LM4. 0: 1. Simulation characteristics with prescribed SSTs. *Journal of Advances in Modeling Earth Systems*, *10*, 691–734. https://doi.org/10.1002/2017MS001208