Downloaded from orbit.dtu.dk on: Apr 27, 2024

DTU Library

Comparison of data-driven methods for linking extreme precipitation events to local and large-scale meteorological variables

Antoniadou, Nafsika; Sørup, Hjalte Jomo Danielsen; Pedersen, Jonas Wied; Gregersen, Ida Bülow; Schmith, Torben; Arnbjerg-Nielsen, Karsten

Published in:
Stochastic Environmental Research and Risk Assessment

Link to article, DOI:
10.1007/s00477-023-02511-3

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Antoniadou, N., Sørup, H. J. D., Pedersen, J. W., Gregersen, I. B., Schmith, T., & Arnbjerg-Nielsen, K. (2023). Comparison of data-driven methods for linking extreme precipitation events to local and large-scale meteorological variables. *Stochastic Environmental Research and Risk Assessment*, *37*, 4337-4357. https://doi.org/10.1007/s00477-023-02511-3

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Stochastic Environmental Research and Risk Assessment (2023) 37:4337–4357
https://doi.org/10.1007/s00477-023-02511-3

ORIGINAL PAPER

Comparison of data-driven methods for linking extreme precipitation events to local and large-scale meteorological variables

Nafsika Antoniadou[1,2] · Hjalte Jomo Danielsen Sørup[1] · Jonas Wied Pedersen[1,2] · Ida Bülow Gregersen[3] · Torben Schmith[2] · Karsten Arnbjerg-Nielsen[1]

Accepted: 28 June 2023 / Published online: 13 July 2023
© The Author(s) 2023

Abstract
Extreme precipitation events can lead to severe negative consequences for society, the economy, and the environment. It is therefore crucial to understand when such events occur. In the literature, there are a vast number of methods for analyzing their connection to meteorological drivers. However, there has been recent interest in using machine learning methods instead of classic statistical models. While a few studies in climate research have compared the performance of these two approaches, their conclusions are inconsistent. To determine whether an extreme event occurred locally, we trained models using logistic regression and three commonly used supervised machine learning algorithms tailored for discrete outcomes: random forests, neural networks, and support vector machines. We used five explanatory variables (geopotential height at 500 hPa, convective available potential energy, total column water, sea surface temperature, and air surface temperature) from ERA5, and local data from the Danish Meteorological Institute. During the variable selection process, we found that convective available potential energy has the strongest relationship with extreme events. Our results showed that logistic regression performs similarly to more complex machine learning algorithms regarding discrimination as measured by the area under the receiver operating characteristic curve (ROC AUC) and other performance metrics specialized for unbalanced datasets. Specifically, the ROC AUC for logistic regression was 0.86, while the best-performing machine learning algorithm achieved a ROC AUC of 0.87. This study emphasizes the value of comparing machine learning and classical regression modeling, especially when employing a limited set of well-established explanatory variables.

Keywords Extreme precipitation · Meteorological drivers · Machine learning · Logistic regression · ROC curve

1 Introduction

Extreme precipitation events can cause significant damage, especially in densely populated regions (Jonkman 2005). An example was in July 2011, when a cloudburst hit the capital of Denmark, Copenhagen, leading to widespread flooding and substantial harm to society, the environment, and the economy (Ziersen et al. 2017). Extreme precipitation events are caused by specific atmospheric processes and can be correlated to both local and large-scale meteorological drivers. Knowledge about these mechanisms will improve our understanding of the local physical climate and our ability to infer under what conditions extreme precipitation events occur and thus enable better predictions of occurrences in the future. Numerous studies have related extreme rainfall to meteorological drivers. North Atlantic Oscillation (NAO) was linked to the variability of the Mediterranean precipitation (Xoplaki et al. 2004), to the occurrence and intensity of extreme precipitation events over northeast Spain (Vicente-Serrano et al. 2009), and to the change in the European winter precipitation and other extremes (Scaife et al. 2008). The East Atlantic Pattern (EA) was related to the convective anomalies in the tropical Atlantic (Maidens et al. 2021) and to the spatial and temporal changes in the frequency of extreme rainfall events over Denmark (Gregersen et al.

✉ Nafsika Antoniadou
nafan@dtu.dk

1 Department of Environmental and Resource Engineering, Climate and Monitoring, Technical University of Denmark, Lyngby, Denmark

2 National Centre for Climate Research, Danish Meteorological Institute, Copenhagen, Denmark

3 Department of Climate Adaptation and Green Infrastructure, Rambøll Denmark A/S, Copenhagen, Denmark

2013a). Extreme precipitation events have been characterized using Convective Available Potential Energy (CAPE) and Dew-Point temperature in the eastern United States (Lepore et al. 2015) and South-Central Andes (Ramezani Ziarani et al. 2019). Furthermore, previous research has related extreme rainfall events with humidity-related variables over the Mediterranean (Hertig et al. 2014; Hertig and Jacobeit 2013) and sea surface temperature (SST) in tropical regions (Dittus et al. 2018).

Many different methods can be used to develop models that explore these relationships. Classic statistical models such as linear (Li and Wang 2018) or logistic (Chan et al. 2018) regression (LR) are commonly applied depending on whether the outcomes are continuous or binary. Regression models are based on theory and explicit assumptions and benefit from domain knowledge for model specification providing a clear framework for understanding the relationships between explanatory variables (Hastie et al. 2009).

Recently the use of machine learning (ML) algorithms is becoming more widespread as an alternative approach for classification and prediction. ML is a subfield of artificial intelligence based on non-linear algorithms adapting and learning from data (Mitchell 1997). These algorithms can process vast amounts of multidimensional data such as reanalysis, satellite, or radar data. ML has been used in predicting extreme rainfall intensities (Davenport and Diffenbaugh 2021; Lee et al. 2012) and for rain/no-rain classification (Liu et al. 2001; Shi 2020).

The differences between ML and classic regression have been extensively explored in the literature (Breiman 2001a). For example, ML automatically includes non-linear associations and interaction terms, whereas for regression methods, they must be specified by the user (Boulesteix and Schmid 2014). Because of this adaptability, ML is claimed to offer superior predictive performance relative to traditional statistical modeling and better handling of a greater number of explanatory variables (Deo and Nallamothu 2016). However, as a downside of this flexibility, ML algorithms tend to overfit the data used for training, which must be compensated for by penalization of the complexity of identified models. In scenarios where data is limited, feature engineering, and feature selection becomes even more important when using ML models to ensure optimal performance and mitigate the risk of overfitting (Chen et al. 2020; Guth and Sapsis 2019).

In this study, the term "prediction" refers specifically to the outcomes generated by statistical and ML models, rather than a forecast of the future state of variable (such as a weather forecast).

Although many studies in other fields (e.g., health sciences) compare the performance of classic statistical models to different ML algorithms, there are only a few within climate sciences. Wei et al. (2020) showed that a decision tree performs better than LR for extreme rainfall event classification. Meyer et al. (2016) proved that a Neural Network (NNET) is a more suitable algorithm for satellite-based rainfall retrievals than Random Forest (RF) and Support Vector Machine (SVM), but LR was not part of the comparison. Lastly, Moon et al. (2019) suggested using LR as an effective early warning system for very short-term heavy rainfall in South Korea instead of ML models.

Even the most sophisticated machine learning algorithms rely on the quality of data. In case of low-quality input data, the reliability of the results will be compromised (Budach et al. 2022). Extreme events are very localized, so choosing a densely monitored study area with rain gauges, such as Copenhagen (Thomassen et al. 2022), is crucial for accurately capturing these events.

The novelty of the present work is two-fold: (i) To identify relationships between meteorological explanatory variables selected based on a priori domain knowledge and extreme events in the densely monitored Copenhagen area. This includes considering both variables previously linked to extreme precipitation events in various regions around the globe and local variables that capture the unique characteristics of the Copenhagen area. By examining the importance of these two types of variables, we aim that our findings can gain valuable insights into the specific drivers of extreme precipitation occurrences in this region and serve as a foundation for extending our analysis to other regions. (ii) To systematically compare the explanatory performance of classification models developed using traditional LR and three ML models and assess the similarities and differences of the most influencing explanatory variables between models. The hypothesis is that traditional LR would result in the lowest performance. In summary, we address the following research questions:

1. Which meteorological drivers can explain local extreme rainfall events in a densely monitored region?
2. What is the relative importance of the drivers across different statistical models?
3. Do ML models lead to improved performance compared to traditional statistical modeling?

2 Data and study area

2.1 Precipitation

Extreme precipitation events are often very localized (< 10 km horizontal scale), and a dense network of rain gauges is therefore needed to record extreme events when they occur. At the same time, to link explanatory variables

with the occurrence of extreme precipitation events in a robust way requires long time series of observations. In this study, the precipitation data set consists of hourly observations for the period 1979–2020 from 15 gauges located in the Copenhagen area. The geographical location of the stations appears in Fig. 1.

The gauges are part of a national network run by the Water Pollution Committee of The Society of Danish Engineers and have a measurement resolution of 0.2 mm (Gregersen et al. 2013a). The gauges have an average uptime of 95.8%, and data has been quality controlled by the Danish Meteorological Institute (DMI). This study only considers the five months of May to September each year, which define the main season of convective rainfall extremes in Denmark (Pedersen et al. 2012).

2.2 Extreme precipitation explanatory variables

This study employs explanatory variables that relate to the physics of convective rainfall events and have been identified in previous research as influential factors for extreme precipitation events. Additionally, local predictors that capture the unique characteristics of the study area are also investigated. The chosen meteorological variables from various sources utilized as explanatory variables over different spatial domains (Fig. 1) are:

1. Observed Surface air temperature (SAT) data for 1979–2020 from one station in the middle of the island Zealand. The temporal resolution changes over time, from 3-hourly observations in the first years to hourly observations in at least the last 20 years. The choice of this variable is motivated by the experience of local meteorologists that extreme precipitation events occur on days with high inland afternoon temperatures when convection may be released. The data were quality controlled by DMI. More information about the observation protocols can be found in the supplementary material.
2. The daily-mean sea surface temperature (SST) for the North Sea and Baltic Sea from the Copernicus Marine Service Information (DMI 2015) for the period 1982–2020. This is motivated by the idea that SSTs could influence convection if the air mass passes over the sea.
3. The rest of the extreme precipitation explanatory variables have been extracted from ERA5, the fifth-generation global reanalysis product by European

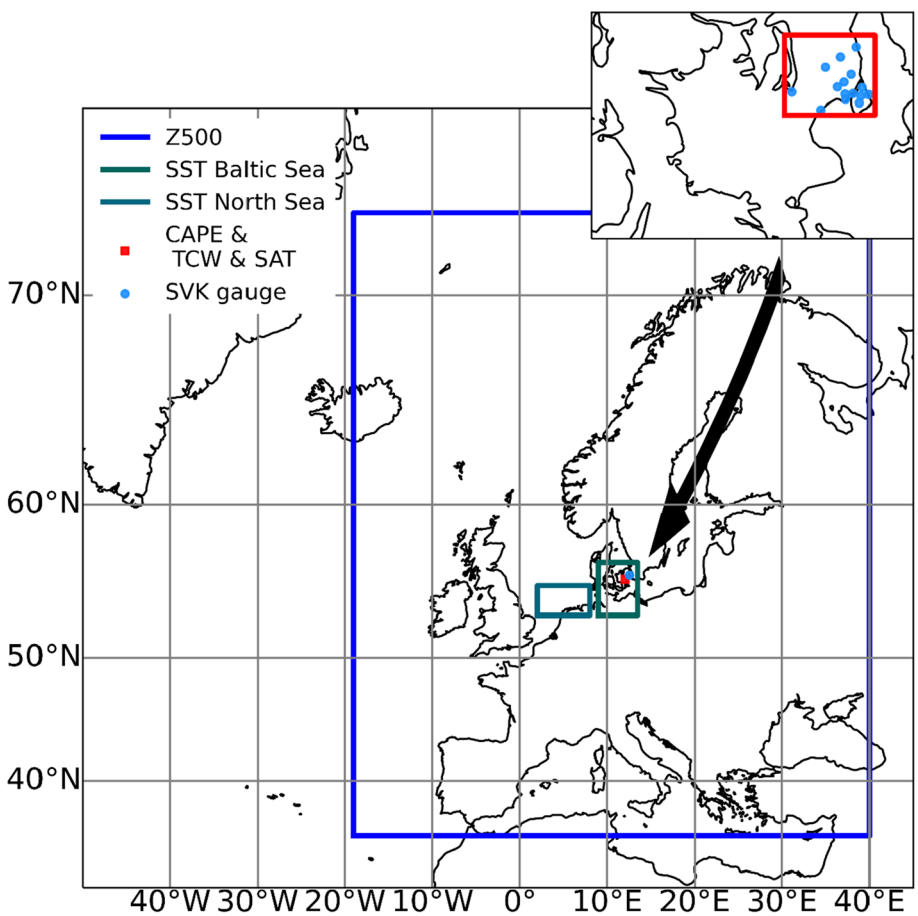


Fig. 1 Overview of the geographical data domains of the explanatory variables and the location of the rain gauge stations in the subplot. There are a total of fifteen stations in an area of 525.50 km^2 (12–12.65 W and 55.57–55.84 N)

Centre for Medium-range Weather Forecasts (ECMWF) (Hersbach et al. 2020) for the period 1979–2020. The data are generated at hourly time steps with a spatial resolution of roughly 30 km × 30 km. The selected explanatory variables are:

- Geopotential height at 500 hPa (Z500), which determines flow strength and direction at 500 hPa, which is usually considered the steering level for meso-scale weather systems.
- Convective available potential energy (CAPE) as a measure of atmospheric instability and the potential for convection due to the vertical temperature gradients and humidity in combination (further details in supplementary material).
- Total Column Water (TCW) as a measure of the maximum amount of water available for precipitation in case of strong convection (further details in supplementary material).

3 Methodology

3.1 Preprocessing data

3.1.1 Definition of an extreme precipitation event

To develop a classification model for predicting extreme precipitation events, class labels first need to be generated. To ensure the independence of events for each station, an 11-h dry period is required between events (Thomassen et al. 2022). For each station, the most extreme events based on the maximum hourly intensity are sampled using a Peak Over Threshold method (POT) with type II censoring (Coles 2001), resulting in 3 events per year on average (Gregersen et al. 2013b). The date of the event was the date when the maximum intensity was observed. Any day with an extreme recorded in at least one station is labeled as an extreme day. This led to a total of, on average, thirteen days with extremes per year for the region.

3.1.2 Explanatory variables

The purpose of including Z500 in this analysis is to examine whether specific patterns in the large-scale, regional atmospheric flow fields are associated with extreme precipitation events in the case area. The daily climatology of Z500 is determined in each grid point by first applying a 5-day centered moving window and then calculating the simple average of all values corresponding to each day of the year. Calculating the climatology of variables with a 5-day smoothing window is a common technique used in climatological studies to reduce the impact of day-to-day variability and reveal long-term patterns. Then anomalies are derived by subtracting the daily climatology from each corresponding day in the selected period, which leads to a regional 2D field of Z500 anomalies for each day. Including Z500 anomaly values from all grid cells in the 2D field would be a huge increase in the number of input predictors for the models. It is therefore common practice in studies like this one to perform Principal Component Analysis (PCA) on the Z500 anomalies as a dimensionality reduction technique (Mastrantonas et al. 2021; Merino et al. 2019; Storch and Zwiers 1984). This way of processing Z500 fields is also referred to as EOF (Empirical Orthogonal Functions) analysis in the field of climatology (Sun and Wang 2018; Yang et al. 2013). PCA enables the representation of the data in a lower-dimensional space, with significantly fewer variables than the original dataset while ensuring independence among these variables. We weigh the data by the square root of the cosine of the latitude to provide equal-area weighting in the covariance matrix. We retain the first Principal Components (PCs), which explain at least 90% of the total variability in the anomaly field. Combining with a scree plot as a graphical representation can provide an adequate validation of the 90% criterion. During exploratory analysis, we also analyzed larger domains than shown in Fig. 1 (i.e., North Atlantic), but this only resulted in weaker ability to explain occurrences of extreme events. Moreover, we followed the domain size recommendations by Chen and Wang (2014) to optimize the effectiveness of PCA.

We then performed K-means clustering as suggested by (Mastrantonas et al. 2021) on the retained PCs, but the performance of the classification models was not improved compared to just including the retained principal components, so for all the analysis hereafter, we use the retained PCs. See Supplementary Material for details.

For CAPE and TCW we are interested in the daily means at the location of the case study, and therefore the data were bilinearly interpolated from the original ERA5 grid cells to the exact geographical position of Copenhagen. For SAT, we extracted the daily maximum (SATmax), the daily difference between the maximum and the minimum (SATamplitude), and the daily maximum of the previous day (SATlag1). Lastly, for SST, the mean, the maximum, and the previous day's maximum value of the North Sea and Baltic Sea (Fig. 1) were tested as individual explanatory variables and as their combinations. An overview of the preprocessing methods can be found in Table 1.

3.2 Classification models

We use four algorithms that can classify a discrete outcome based on continuous input: LR as a traditional statistical

Table 1 Overview of the preprocessing methods for each explanatory variable

Explanatory variable	Preprocessing analysis
500 hPa geopotential height (Z500)	1) PCA
	2) PCA + k-Means
Convective available potential energy (CAPE)	1) Daily average + bilinear interpolation
Surface air Temperature (SAT)	1) Daily max
	2) Daily amplitude
	3) Daily max lag one day
Total Column Water (TCW)	1) Daily average + bilinear interpolation
Sea surface temperature (SST)	1) Daily mean of North/Baltic/ North & Baltic Sea
	2) Daily max of North/Baltic/ North & Baltic Sea
	3) Daily max lag one day of North/Baltic/ North & Baltic Sea

modeling approach and RF, NNET, and SVM as ML algorithms. All ML algorithms have one or more hyper-parameters that control how well the model fits the data, and the optimal values for these parameters can vary from dataset to dataset. We do not impose pre-determined interactions on the ML models to allow them to leverage their inherent capabilities in handling variable interactions. This approach enables fair comparisons and allows each model to utilize variables and their interactions in the manner that best complements its underlying algorithm.

3.2.1 Logistic regression models

LR analysis is the most frequently used modeling approach for analyzing dichotomous response variables (i.e., occurrence or non-occurrence of an extreme event). It belongs to the family of generalized linear models (GLM). In a GLM, the three building blocks are (Lindsey 2000): a random component, a systematic component, and a link function. In LR, the random component Y follows a binomial distribution and can be represented by the model:

$$G(E(Y)) = X\beta^T + e \tag{1}$$

where G is a link function, X is the design matrix of n systematic components (explanatory variables), β^T is a vector of coefficients, and e is the residual vector term.

In this study, we use the logit link function (Cox et al. 1999):

$$G(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) \tag{2}$$

which monotonically maps the domain $(-\propto, +\propto)$ to (0,1) where π is the probability of an extreme event and $\frac{\pi}{1-\pi}$ is the corresponding odds of an event being extreme. The estimate of π is then given by:

$$\pi = \frac{\exp\left(X\beta^T\right)}{1+\exp\left(X\beta^T\right)} \tag{3}$$

The coefficients β_i, $i = 1, \ldots, p$ are estimated using the maximum likelihood method (ML). The significance of the individual coefficients is assessed with the Wald statistic, which is simply the ratio:

$$Z = \frac{\hat{\beta}_i}{s.e.\left(\hat{\beta}_i\right)} \tag{4}$$

which can be used to test if $\beta_i = 0$. The standard normal distribution is used to determine the p value of the test, and the confidence intervals are given by:

$$\hat{\beta}_i \pm z_{1-a/2} s.e.\left(\hat{\beta}_i\right) \tag{5}$$

In addition to estimating the probability of an extreme event, LR also provides a measure of association between the response variable and the explanatory variables in the form of odds ratios. The odds ratio represents the change in the odds of the response for a one-unit increase in the explanatory variable, holding all other variables constant. In LR, the odds ratio is calculated as the exponentiated coefficient of the explanatory variable. An odds ratio with a value equal to one indicates no association between the explanatory variable and the response, while a value higher/lower than one indicates a positive/negative association.

Understanding the strength of the relationships between the explanatory variables in a regression model is crucial since it may affect the reliability of the model's estimated coefficients. Variance Inflation Factors (VIF) is one way to assess the potency of these connections (VIF). The VIF (O'brien 2007) is a statistical measure that quantifies the degree to which the variance of the estimated regression coefficient β_i, increases due to the presence of multi-collinearity among the explanatory variables in the model. The calculation of VIF involves regressing each explanatory variable on all the other variables in the model, and the VIF is for the ith explanatory variable is determined as:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{6}$$

where R_i^2 is the R^2-value obtained by regressing the ith explanatory variable on the remaining variables. A VIF value of 1 indicates no collinearity, while values greater than 1 indicate an increasing degree of multicollinearity.

Since we are analyzing time series data, it is important to examine the possibility of serial correlation in the regression residuals, which means the deviation from the expected characteristics of white noise residuals in LR. Serial correlation does not prevent precise predictions of the response within the model's scope, but the standard errors of the estimated odds ratios will, in general, be underestimated. To evaluate the impact of serial correlation, we examine the significance of the regression odds ratios for the model fitted to all data versus to a thinned data set that sub-samples every third and every fifth day, respectively. Thinning the data this way reduces the serial correlation in the response and, therefore, likely also in the regression residuals. By doing so, we aim to examine the robustness of our results to different levels of serial correlation. The choice of this thinning scheme is based on the partial autocorrelation function (See Supplementary Material).

Lastly we did experiments with modeling the explanatory variables with restricted cubic splines (Gauthier et al. 2020) to address potential nonlinearity, but the use of splinesdid not lead to any significant increase in the performance of the final model.

3.2.2 Random forest

Since its introduction in 2001 (Breiman 2001b), the RF supervised machine learning algorithm for classification has seen a significant increase in popularity. RF is an ensemble learning method consisting of multiple decision trees that result in reduced variance compared to single decision trees (James et al. 2021). Each of the trees fit in a different bootstrap sample of the original dataset (bagging) to increase the diversity in the decision trees. To further decorrelate the decision trees, RF randomly re-samples the explanatory variables at each split. In contrast to LR, RF is not restricted by the assumption of independence between explanatory variables, captures the interactions between explanatory variables without specifying them, and can model complicated non-linear effects. One of the primary hyperparameters is the number of trees. It is often regarded as best practice to grow as many trees as computationally possible.

We use the Gini impurity (Nembrini et al. 2018) to determine how the explanatory variables of the dataset should optimally split nodes when training a decision tree:

$$Gini = 1 - \sum_{r=1}^{2} (\pi_r)^2 \tag{7}$$

where r represents the index of the classes in the dataset.

3.2.3 Support vector machine

A SVM is a supervised ML model for classification (Cristianini and Shawe-Taylor 2000). The objective of SVM is to find a hyperplane in the n-dimensional space (n = the number of explanatory variables) that distinctly classifies the data. The hyperplane is defined as the set of points x satisfying:

$$w \cdot \mathrm{x} + \mathrm{b} = 0 \tag{8}$$

The vector w and scalar b for the best hyperplane are determined by an optimization procedure that maximizes the margin between two classes in the n-dimensional space.

We use a soft margin technique that allows for a number of misclassified cases. This number is controlled by a hyperparameter (cost) which imposes a penalty on the model for making an error. Moreover, since the classification problem is non-linear, we use a kernel function that returns the dot product of the transformed vectors in the higher dimensional space. We used the Gaussian radial basis function

$$G(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) \tag{9}$$

where σ is the bandwidth of the kernel function hyperparameter and x_i and x_j represent two different data points from the dataset.

The SVM model performs poorly on imbalanced datasets (Fernández 2018), so we use an extension to the algorithm in order to increase its performance: we create weights for our data samples such that each sample is weighted according to its corresponding class (extreme/non-extreme) size. Samples of bigger classes will be assigned smaller weights and vice versa.

3.2.4 Feed-Forward neural network

NNET is a family of ML algorithms that use one or more layers of nodes (also known as neurons) coupled by non-linear functions with adjustable parameters (weights) to map inputs to outputs (Bishop 2006). Network training aims to find a set of weights that minimizes the difference between the NNET output and the training labels. The loss function in classification models is usually cross-entropy

(Eq. 4.90 in Bishop 2006). The loss function is then minimized using one of the gradient descent techniques, whereby a step is made in the direction of the steepest fall (the negative of the gradient), with the size of the step regulated by the learning rate. However, even though NNETs may approximate any smooth non-linear function and allow for collinearity among the explanatory variables, they are challenging to understand and can also be computationally expensive to train.

3.3 Model tuning and validation

The data set is divided into training/validation data (80%) and testing data (20%). Within the training/validation subset, a tenfold cross-validation technique is applied to select the best model (Kohavi 1995). Training data are used for model training and fitting internal model parameters, while validation data are used for tuning model hyperparameters and variable selection. In the tenfold cross-validation process, the training/validation data are randomly partitioned into ten folds of approximately equal number of years. Following that approach, there are samples with specific physical properties for generating extreme precipitation events in all subsets. At each iteration, one of the folds is chosen as the validation set, while the remaining folds are used for the training set. The model is fitted in the training set, and the performance metric is computed based on the validation set. Finally, the performances are averaged across the ten iterations, and for each algorithm, the model that performs best across all the iterations is selected as the final model. This final model is then evaluated on the independent testing data, which was initially set aside.

3.4 Performance metrics

Given the rarity of extreme events, the data are characterized by a very low fraction of days with extremes. In light of this, the dataset is imbalanced, so using a traditional performance metric such as accuracy (ACC) could be misleading since it can be maximized by simply predicting the majority class and thus omitting the minority class (Liu et al. 2009). ACC is defined as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and TN = False Negatives, which are the four possible outcomes of binary predictions.

The receiver operating characteristic curve (ROC) (Mason and Graham 2002) is a commonly used validation tool for classification problems and consists of a two-dimensional graph that shows the true positive rate (TPR):

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

versus the false positive rate (FPR):

$$\mathrm{FPR} = 1 - \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{12}$$

for all possible probability thresholds.

A random model will produce the diagonal line as its ROC curve, and a perfect model will have a ROC curve composed of the left and upper boundary lines. The "steepness" of ROC curves is hence important since maximizing the benefits (TPR) and minimizing the costs (FPR) is ideal. ROC AUC measures the area underneath the entire ROC curve and is the measure of the ability of a model to distinguish between two classes. It provides the total performance measure across all potential classification thresholds and varies between zero and one. The higher the ROC AUC, the better the model is at making classifications. The Mann–Whitney U test (Wilks 2011) is applied to estimate whether the model performs statistically better than a random model in terms of ROC AUC. The DeLong test (DeLong et al. 1988) was used to make pairwise comparisons in ROC AUC between LR and ML models. Statistical significance is defined at $\alpha = 0.05$.

On the other hand, the area under the Precision-Recall (PR) curve (Saito and Rehmsmeier 2015) summarizes the trade-off between precision and TPR at all possible probability thresholds, taking into account the imbalance in the dataset. Precision is the fraction of TP predictions among the positive predictions made by the model:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

In imbalanced datasets, it is possible for a classifier to achieve a high ROC AUC by making a large number of false positive predictions, especially when the positive class is rare. In this case, the PR AUC will be much lower, reflecting the low quality of the positive predictions. Therefore, it is important to consider both the ROC AUC and the PR AUC when evaluating the performance of a binary classifier in imbalanced datasets, as they capture different aspects of the classifier's behavior (Davis and Goadrich 2006).

To assess if the models are well-calibrated, we use the Brier score. The score is given by:

$$Brier\ score = \sum_{t=1}^{n} (y_t - o_t)^2 \tag{14}$$

where y_t is the predicted value, o_t is the true value, and n is the number of observations. The Brier score also has shortcomings for imbalanced datasets (Benedetti 2010), but it can be used as a relative measure for model comparison.

We will use ROC AUC as the primary measure of model performance, and the secondary measures of performance are the PR AUC, the Brier score, and *ACC* together with *TPR* and *FPR* at the optimal threshold, which is the threshold that maximizes the sum of *TPR* and $1 - FPR$.

In this study the variable selection on ML models is done in same way as with the regression model. This decision has been made to establish a consistent and comparable framework for evaluating the different modeling approaches. By using the same set of variables across all models, we could directly compare the performance of the ML models against the LR model, highlighting the potential benefits of ML methods in terms of capturing complex interactions.

3.5 Variable importance

The contribution of the different explanatory variables to the overall performance was quantified for all models. For LR, we study the variable importance of the explanatory variables in explaining extreme events by using the difference in the deviance of the full model (Pawitan 2014) and a model without the explanatory variable whose importance we want to assess. The deviance test statistic,D^*, is given as:

$$D^* = -2(\ell(\beta^0) - \ell(\beta)) \tag{15}$$

where (β^0) is the log-likelihood of the reduced model, and (β) is the log-likelihood of the full model. This test statistic has a chi-square distribution with 1 degree of freedom. If H_0 is rejected, there is evidence that the left-out explanatory variable contributes significantly to the prediction of the outcome.

For RF and SVM, the variable importance is the decreased ROC AUC after the permutation of the variable series (Breiman 2001b). The idea is that if the values of an important variable are permuted, keeping all other variables the same, the performance would degrade. An explanatory variable is important if permuting its values decreases the model ROC AUC relative to the other variables and unimportant if permuting its values keeps the ROC AUC almost unchanged. There can be cases where permuting a variable with very little explanatory power can cause an increase in ROC AUC due to random noise. This will end up with negative importance scores equivalent to zero importance. Since the model includes only continuous explanatory variables, there is no bias in the permutation importance measure (Strobl et al. 2007). We repeat the same process five times to increase the stability of our estimates.

Finally, for NNET, the variable importance is assessed by the "weights" method (Gevrey et al. 2003). The method entails decomposing the hidden-layer connection weights of each output neuron into components associated with each input neuron.

In addition to the previously described methods for assessing variable importance in our study, we acknowledge that even if these methods are widely used, they might not capture complex interactions and dependencies among variables in the model. To overcome these limitations, we have used the SHAP (Shapley Additive Explanations) method (Lundberg and Lee 2017) as a complementary analysis in our study. SHAP leverages game theory principles to identify the relative importance value of each explanatory variable, considering all possible combinations of variables. This approach allows us to capture both linear and non-linear effects, as well as interactions between variables, providing a more accurate and interpretable measure of variable importance. SHAP values offer insights not only into the global importance of variables but also into how variables contribute to the prediction of each individual instance. The global importance can be obtained as the average of absolute SHAP values for each explanatory variable. Moreover, the SHAP method is agnostic, as it can be applied to a wide range of models, including LR, RF, SVM, and NNET. This property enables us to assess variable importance consistently across different models, improving the comparability and generalizability of our results.

3.6 Computation

The Principal Component analysis was conducted with the Python package eofs (Dawson 2016). All the classification models have been developed using the R package caret (Kuhn 2008). We used the R package vip (Greenwell and Boehmke 2020) to quantify the variable importance of RF and SVM and kernelshap R package for finding the SHAP values (Mayer and Watson 2023). ROC AUC curve comparisons were conducted using the R package pROC (Robin et al. 2011).

4 Results

4.1 Resolving geopotential height into principal components

Figure 2 shows the eigenvalues for each PC. We can see that the elbow flattens out at nine PCs. In addition, the first nine PCs account for 90% of the total variability for Z500. Therefore, we regard the first nine PCs as adequate to describe the spatiotemporal variation of the geopotential height.

Figure 3 shows the spatial patterns associated with, and the proportion of the overall variance explained by each of

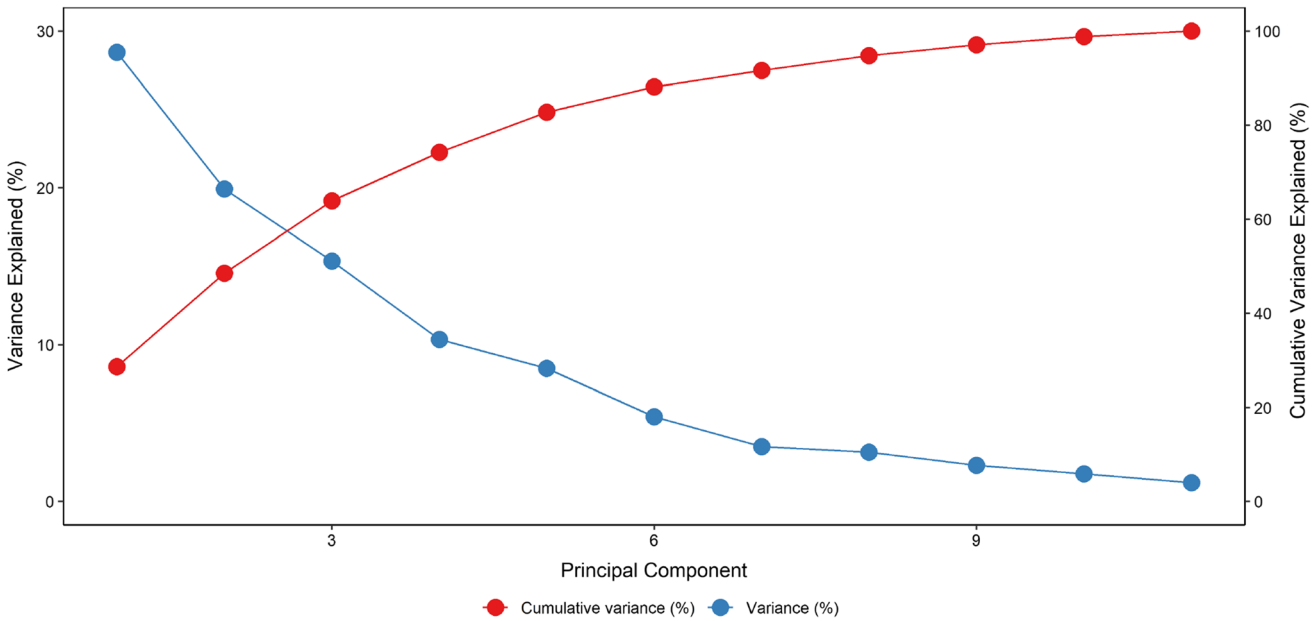


Fig. 2 Scree Plot showing variance explained by Principal Component Analysis on Z500, with cumulative variance on the secondary y-axis

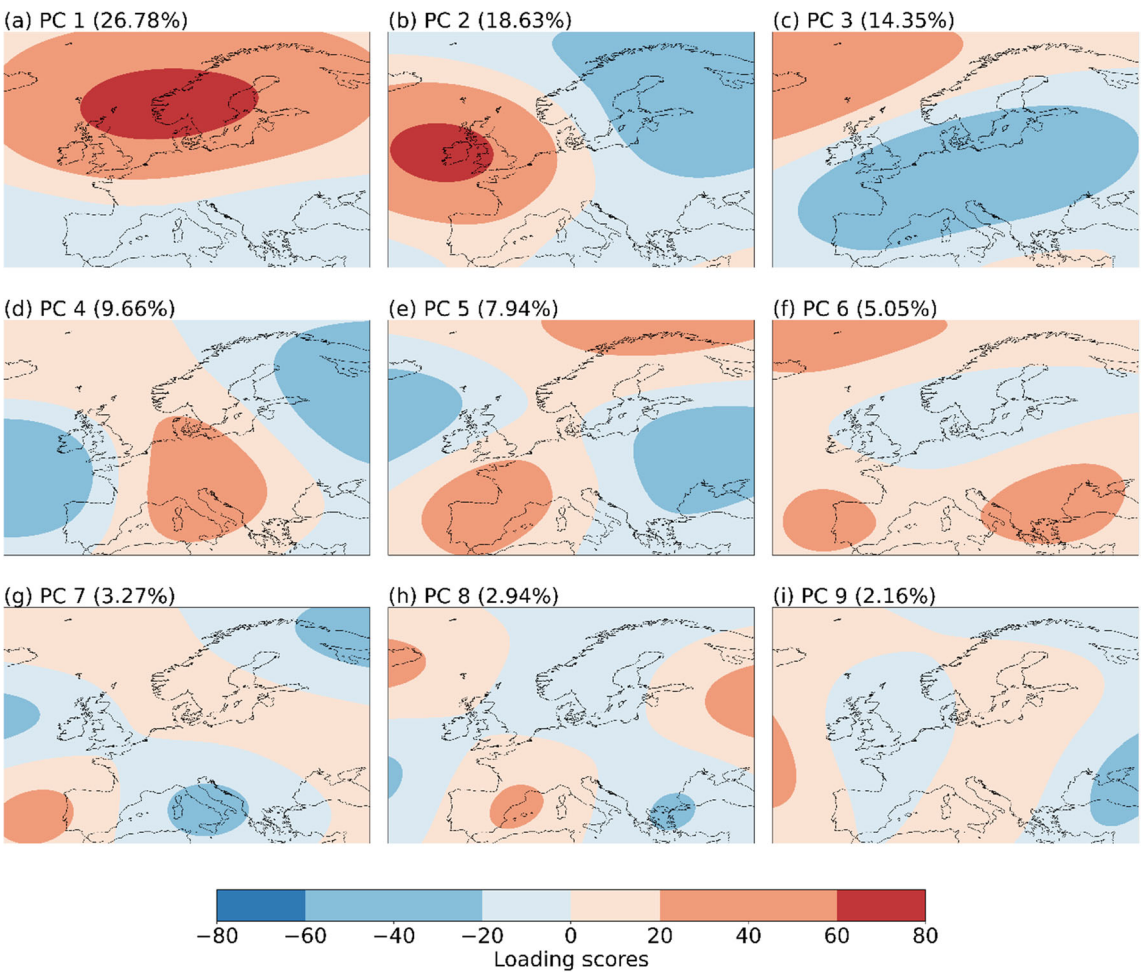


Fig. 3 The first nine principal components of daily anomalies at a geopotential height at 500 hPa

the PCs. The first two PCs, which collectively describe 45% of the variance, exhibit spatial structures that are very similar to well-known circulation patterns over Europe. PC1 corresponds to the summer North Atlantic Oscillation (SNAO) described in (Folland et al. 2009). The SNAO is a dipole pattern with nodes of opposite polarity over central Scandinavia and over Greenland. The positive (negative) phase of SNAO results in warm (cool) and dry (wet) summers in Scandinavia. PC2 corresponds to the East Atlantic pattern (EA), which, unlike the SNAO, exists throughout the entire year (Wulff et al. 2017). Here one node is located west of England, and the other (of opposite polarity) over Eastern Europe and the Mediterranean.

4.2 Variable selection

For variable selection, we used the data from 1982 to 2020 to have a complete record of all explanatory variables identified in Sect. 2.2. We first investigate dependencies between explanatory variables, and then we fit and train LR and ML models to different combinations of the explanatory variables. For SAT, the SATmax, and for SST, the Daily max lag of one day of North & Baltic Sea manipulations gave the best results. If CAPE and SATmax were in the model, then their interaction is added in the LR models as an explanatory variable. It is important to note that during the variable selection process on ML models, we found that the chosen explanatory variables are the same as those employed by LR. Therefore, this section presents the variable selection results for only the LR model. For ML models, see supplementary material.

It can be seen from Fig. 4 that all the VIFs except SATmax, SST, and TCW are less than two, which suggests a weak correlation between a given explanatory variable and other variables in the model. To better understand why SATmax, SST, and TCW have higher VIFs, we calculated the matrix of correlations between all possible pairs of explanatory variables. The correlation matrix shows that SATmax is moderately correlated to SST and TCW, and SST is moderately correlated with TCW, which explains why they have higher VIF values. We also reproduce the positive correlation between PC1 and SATmax reported in (Folland et al. 2009). PC2 is a pattern of strong northwesterly flow over Denmark at 500 hPa, and therefore the negative correlation between PC2 and both CAPE and TCW means that unstable and moist air comes with strong southeasterly flow, and this is in accordance with domain knowledge (e.g., Solantie et al. 2006). Furthermore, PC3 has positive correlations with SATmax and TCW. This is not straightforward to explain since PC3 represents a northeasterly flow with its maximum intensity north of Denmark.

Table 2 shows the performance metrics of each explanatory variable independently and the combinations of variables with the best performance. The Brier score is close to zero and almost the same for all combinations of explanatory variables, which means that the models are equally well-calibrated. In the ROC curve analysis, the mean AUC for all models except SATmax and SST alone is above 0.70, which shows that models based on CAPE, PC Z500, or TCW have good discriminative ability. The Mann–Whitney U test also suggests skillful models with $p < 0.05$ (not shown). However, these single-variable models have FPR values around 0.29, which for an imbalanced dataset like the one used in our analysis means a lot of false positives. That also shows up in a poorer PR curve in Fig. 5.

The model combining CAPE, PC Z500, and TCW as explanatory variables outperform the single-variable models. Adding SATmax as an explanatory variable increases the PR AUC only a little, but because the interaction with CAPE is statistically significant (Table 3), we choose to include this variable in the model. On the other hand, adding SST provides no additional performance improvements. It can be seen from Fig. 5 that the models "PC Z500 & CAPE & SATmax & TCW" and "PC Z500 & CAPE & SATmax & TCW & SST" have almost the same PR AUC (the highest). Therefore, PC Z500, CAPE, SATmax, and TCW are considered the most important explanatory variables, and we use them in further evaluation.

Following the model evaluation procedures of Sect. 3.2.1, CAPE, TCW, PC1, PC3, PC4, PC6, the interaction of CAPE and SATmax are found to be statistically significant explanatory variables for the model trained with all and the thinned data (Table 3). We can conclude that for these explanatory variables, the serial correlation of residuals does not affect the confidence intervals to the extent that inference results are at risk of being misinterpreted. The variables CAPE and TCW stand out as having a positive influence on the probability of extreme rainfall; the odds ratios of SATmax are larger than unity, meaning that high SATmax favors extreme precipitation, although the odds ratio is not robust across the subsampled datasets. These findings are in accordance with the arguments we gave in Sect. 2.2 for including these explanatory variables.

The interaction of CAPE and SATmax seems to have a (slightly) negative impact on the probability of an extreme event. In this case, a significant odds ratio of 0.99 for the interaction term suggests that the effect of CAPE on the probability of an extreme event becomes weaker as SATmax increases. Therefore, even though the odds ratios of CAPE and SATmax are both larger than unity, the interaction term decreases the overall effect of CAPE and SATmax on the probability of an extreme event. The interaction term is best understood by examining how the

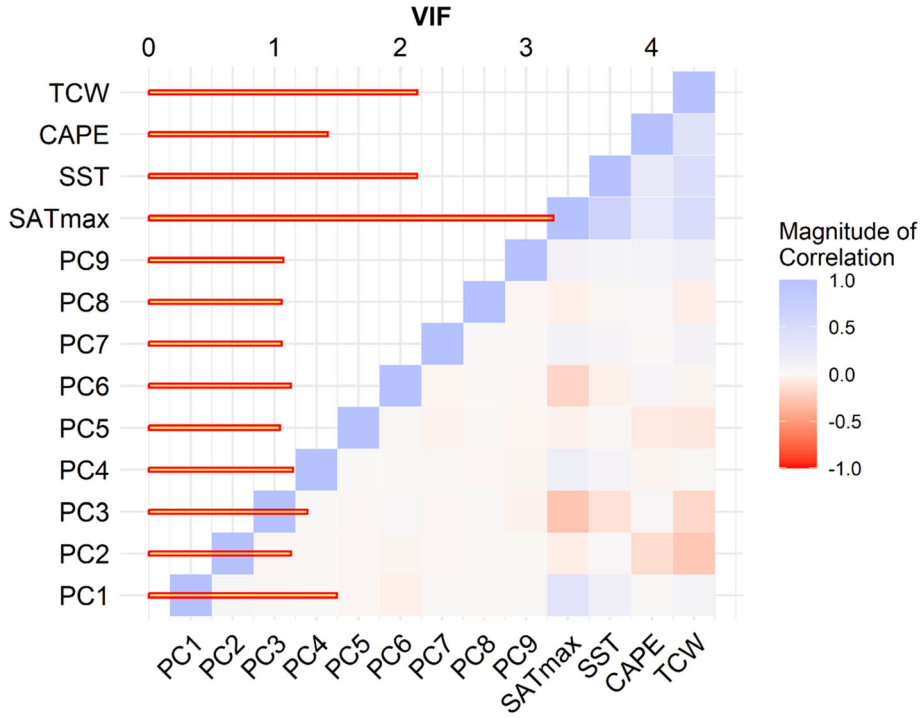


Fig. 4 Correlation matrix depicting correlation of explanatory variables (lower right corner) and VIFs between all explanatory variables (left)

Table 2 Performance metrics for LR models with different (combinations of) explanatory variables for 1982–2020. Data are displayed as mean (Cross-validation values of the first quartile/third quartile)

	ACC	FPR	TPR	ROC AUC	PR AUC	Brier
CAPE	0.72 (0.68–0.76)	0.29 (0.24–0.33)	0.87 (0.84–0.91)	0.84 (0.81–0.87)	0.29 (0.25–0.36)	0.07 (0.06–0.09)
PC Z500	0.62 (0.56–0.7)	0.4 (0.31–0.48)	0.78 (0.69–0.89)	0.72 (0.7–0.74)	0.17 (0.15–0.2)	0.08 (0.07–0.09)
TCW	0.64 (0.56–0.71)	0.37 (0.29–0.48)	0.74 (0.64–0.83)	0.72 (0.71–0.74)	0.19 (0.15–0.24)	0.08 (0.07–0.09)
SATmax	0.42 (0.3–0.55)	0.61 (0.45–0.76)	0.75 (0.58–0.9)	0.53 (0.5–0.57)	0.09 (0.08–0.11)	0.08 (0.07–0.09)
SST	0.52 (0.44–0.65)	0.49 (0.35–0.58)	0.71 (0.64–0.74)	0.59 (0.55–0.62)	0.11 (0.08–0.14)	0.08 (0.07–0.09)
CAPE & TCW	0.73 (0.7–0.79)	0.27 (0.19–0.31)	0.75 (0.69–0.8)	0.79 (0.78–0.8)	0.24 (0.21–0.29)	0.07 (0.07–0.09)
PC Z500 & CAPE & TCW	0.77 (0.73–0.82)	0.24 (0.18–0.28)	0.81 (0.76–0.87)	0.84 (0.83–0.86)	0.33 (0.29–0.39)	0.07 (0.06–0.08)
PC Z500 & CAPE & SATmax & TCW	**0.78 (0.73–0.83)**	0.23 (0.16–0.28)	**0.81 (0.81–0.84)**	**0.86 (0.84–0.87)**	**0.38 (0.34–0.42)**	**0.06 (0.06–0.08)**
PC Z500 & CAPE & SST & TCW	0.77 (0.73–0.81)	**0.23 (0.19–0.28)**	0.8 (0.78–0.87)	0.84 (0.83–0.86)	0.33 (0.29–0.38)	0.07 (0.06–0.08)
PC Z500 & CAPE & SATmax & SST & TCW	**0.78 (0.74–0.82)**	0.22 (0.18–0.27)	0.81 (0.78–0.84)	**0.86 (0.84–0.87)**	**0.38 (0.34–0.42)**	**0.06 (0.06–0.08)**

The best model based on each individual metric score is bold

odds ratio of CAPE in the LR changes as the values of SATmax change (Fig. 6).

Turning to the PCs of Z500, the odds ratio of PC1 is smaller than unity, and the odds ratio of PC3 is larger than unity, and by conferring with Fig. 3, we can interpret that as easterly flows increase the probability of an extreme precipitation event. Similarly, the odds ratio of PC4 being below unity can be interpreted as southeasterly flow increases the probability of extreme precipitation events. This is in agreement with local experience.

The PC6 has a pattern, which implies a stagnant atmosphere over Denmark. Therefore it is surprising that it has a significant odds ratio. That said, the exact shape of the higher-order PC usually is sensitive to the data. This effect is not included in the uncertainty intervals in Table 3, which, therefore, may be underestimated.

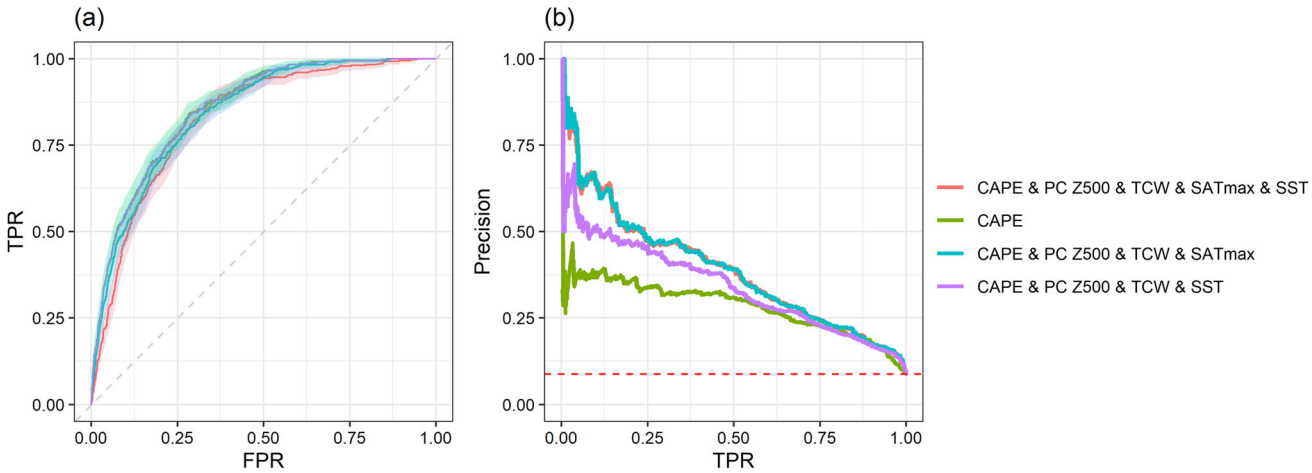


Fig. 5 ROC curves (**a**) for the models with the highest ROC AUC and the corresponding PR curves (**b**) for the period 1982–2020. The dashed lines represent the baseline. The colored bands indicate the 95% CI estimated with stratified bootstrapping

Table 3 The LR odds ratios and the corresponding confidence intervals in parenthesis for all data, and thinned to every third and every fifth day, respectively

Explanatory variable	All data (95% CI)	1/3 Days (95% CI)	1/5 Days (95% CI)
CAPE	**1.04 (1.03–1.05)**	**1.04 (1.03–1.06)**	**1.03 (1.02–1.05)**
TCW	**1.18 (1.15–1.21)**	**1.2 (1.14–1.25)**	***1.17 (1.11–1.24)***
SATmax	1.01 (0.96–1.05)	1.05 (0.97–1.14)	0.99 (0.9–1.1)
PC1	**0.51 (0.43–0.59)**	**0.55 (0.42–0.71)**	**0.58 (0.42–0.79)**
PC2	0.91 (0.8–1.04)	0.99 (0.78–1.24)	0.9 (0.68–1.18)
PC3	**1.47 (1.29–1.67)**	**1.57 (1.25–1.98)**	**1.47 (1.13–1.93)**
PC4	**0.73 (0.64–0.83)**	**0.76 (0.61–0.94)**	**0.69 (0.53–0.9)**
PC5	0.98 (0.87–1.11)	1.05 (0.85–1.29)	1.16 (0.91–1.49)
PC6	**1.66 (1.45–1.89)**	**1.66 (1.33–2.09)**	**1.73 (1.33–2.28)**
PC7	0.93 (0.83–1.05)	0.96 (0.78–1.19)	0.96 (0.75–1.24)
PC8	**1.15 (1.01–1.31)**	1.19 (0.95–1.49)	1.27 (0.97–1.67)
PC9	0.96 (0.85–1.09)	0.98 (0.79–1.21)	0.9 (0.69–1.18)
CAPE* SATmax	**0.999 (0.998–0.999)**	**0.998 (0.998–0.999)**	**0.999 (0.998–1)**

Significant p values are indicated in bold. Significance is evaluated at a 5% level. Confidence intervals are based on the profiled log-likelihood function

4.3 Variable importance

Regarding hyperparameter tuning for RF, an ensemble of 700 trees was found to be sufficient for achieving good performance, with negligible benefits for larger forests. The optimal hyperparameters for each ML algorithm can be found in the supplementary material. We do not impose pre-determined interactions on the ML models to allow them to leverage their inherent capabilities in handling variable interactions. This approach enables fair comparisons and allows each model to utilize variables and their interactions in the manner that best complements its underlying algorithms.

CAPE and TCW obtained the highest importance across all models, with CAPE being the first by RF, NNET, and TCW for LR and SVM. The explanatory variables PC1 and PC6 were consistently rated among the top six variables of the models. A difference was that PC2 was among the top 3 variables in the NNET model, but it had a minor contribution to LR, RF, and SVM. The most notable difference was observed regarding the SATmax, which played no role in LR and a minor role in the RF model, while it was the fourth most important explanatory variable for predicting extreme events in the NNET, SVM model (Fig. 7).

Regarding SHAP values (Fig. 8) CAPE and TCW also obtained the highest importance across all models, with CAPE being also the first by RF, NNET, and TCW by LR and SVM. PC1 was among the top four variables of all the models. SATmax has a larger importance for NNET and SVM compared to LR, and RF, which is in accordance with Fig. 7. In Fig. 8, SATmax is ranked higher for all models except NNET compared to its ranking in Fig. 7.

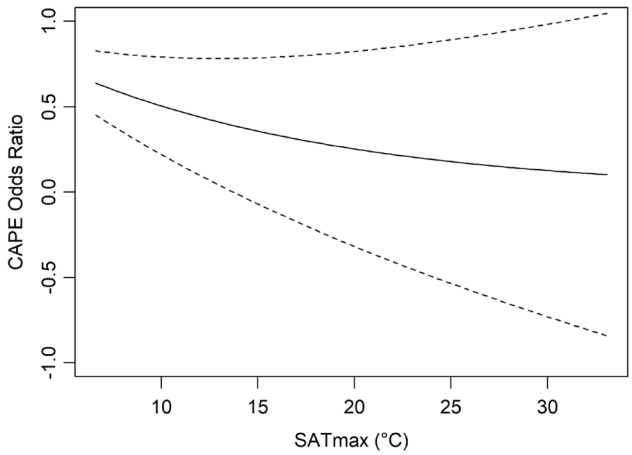


Fig. 6 CAPE odds ratio as a function of SATmax. The solid line is the central estimate; dashed lines are the 95% confidence interval calculated using the Delta Method (Oehlert 1992). The trend indicates that the overall effect of CAPE on extremes decreases with increasing values of SATmax

This difference suggests that SATmax is more important in predicting extreme precipitation events when considering the SHAP values. Figure 8 (right) shows the bee swarm plots where the explanatory variables are not only ordered by their effect on the prediction but also provide insights into how higher and lower values of each variable will affect the outcome. Each dot represents an observation. It can be seen that for all models, higher values of CAPE and TCW have a positive effect on the prediction, while higher values of PC1 have a negative effect. In general, the results for all models in Fig. 8 (right) agree with the LR model coefficients of the statistically significant explanatory variables that are robust across the subsampled datasets (Table 3).

The SHAP interaction values for CAPE and SATmax across all models demonstrate that the importance of CAPE on extreme events decreases as SATmax increases (see supplementary material). This observation is consistent with the findings presented in Fig. 6 where the relationship between the interaction and extreme events was specifically examined within the LR model.

4.4 LR versus ML

In contrast to LR and NNET, ROC AUC decreased for RF and SVM when applied to the test set (Table 4) compared to the training set, but all models still showed high performance (ROC AUC greater than 0.80). In the test dataset, the ML algorithms with the greatest ROC AUC were RF, NNET (AUC = 0.87) (Fig. 9), followed by LR (AUC = 0.86), and SVM (AUC = 0.80). However, the Delong-test results in Table 4 indicated that there were significant differences only between ROC AUC of SVM and that of the other models. For accuracy at the optimal threshold, LR and RF were the best-performing algorithms for predicting extremes events (ACC = 0.75, 0.73), and they had the smallest FPR (FPR = 0.26, 0.28), but NNET had the

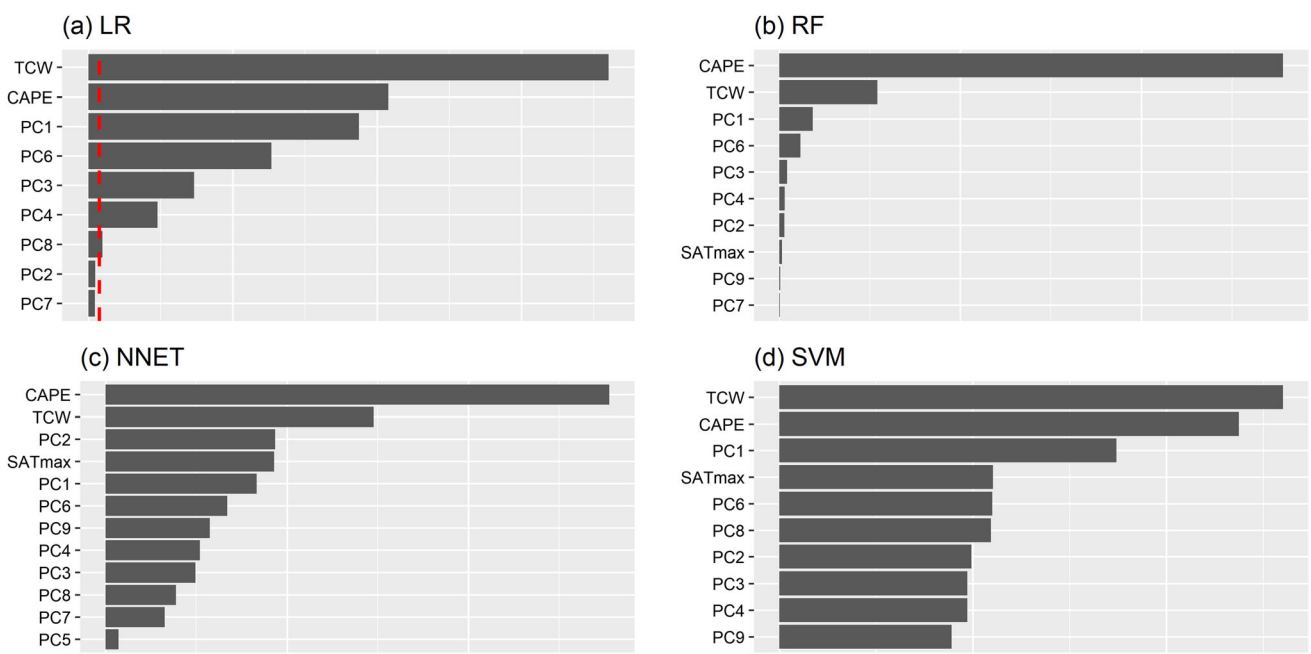
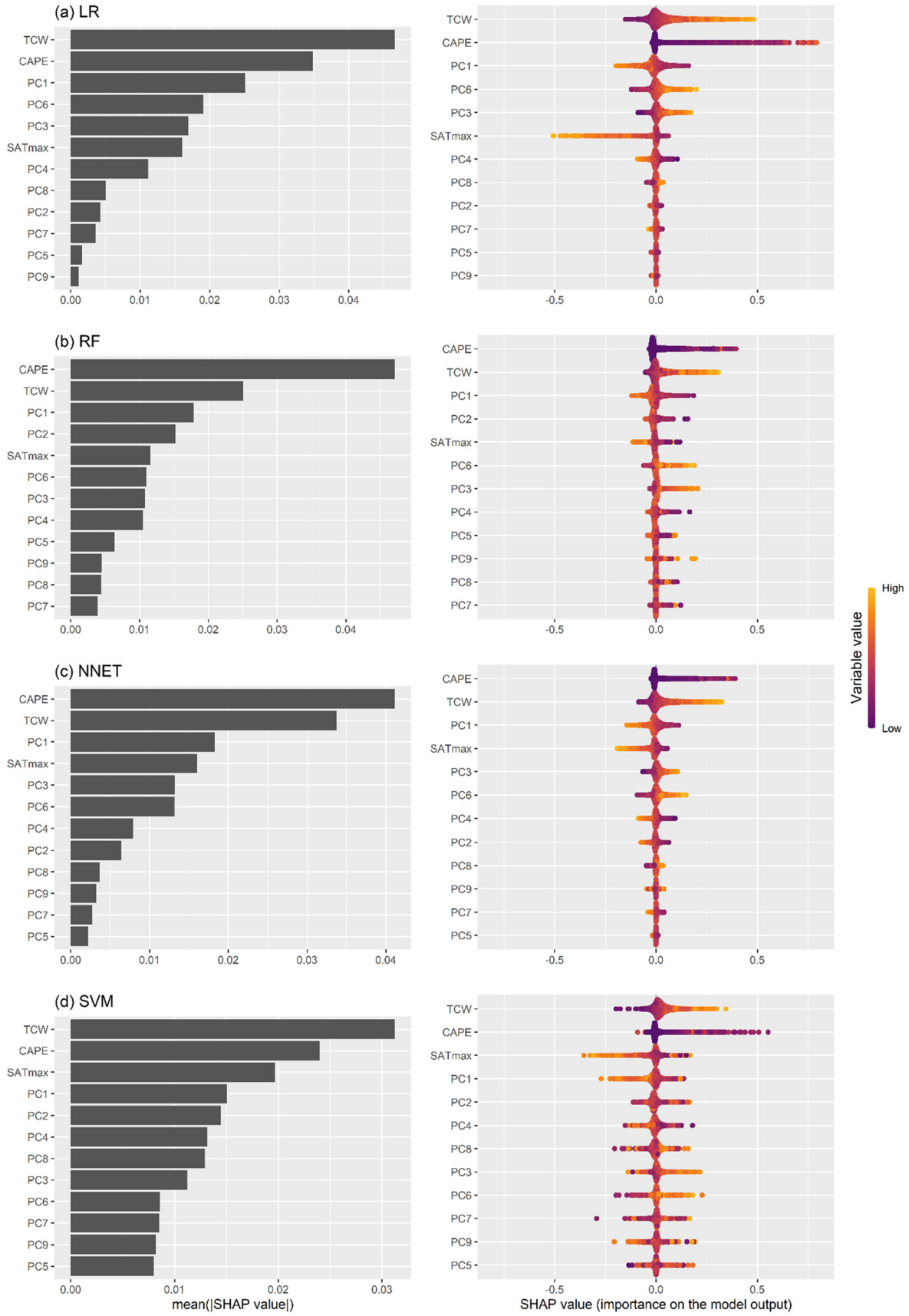


Fig. 7 Importance of explanatory variables in predicting extreme precipitation events. The Type II Chi-square deviance test statistic for LR, permutation-based performance for RF, SVM, and Gevrey Importance for NNET. The dashed line signals statistical significance threshold (p values ≤ 0.05). If a particular explanatory variable is absent from the plot, it indicates that it was considered unimportant by the model


(a) LR
TCW
CAPE
PC1
PC6
PC3
SATmax
PC4
PC8
PC2
PC7
PC5
PC9
0.00
0.01
0.02
0.03
0.04
-0.5
0.0
0.5
(b) RF
CAPE
TCW
PC1
PC2
SATmax
PC6
PC3
PC4
PC5
PC9
PC8
PC7
(c) NNET
CAPE
TCW
PC1
SATmax
PC3
PC6
PC4
PC2
PC8
PC9
PC7
PC5
(d) SVM
TCW
CAPE
SATmax
PC1
PC2
PC4
PC8
PC3
PC6
PC7
PC9
PC5
mean(|SHAP value|)
SHAP value (importance on the model output)
Variable value
High
Low

◀**Fig. 8** In the left: variable importance, evaluated using the mean absolute SHAP values, for all models. In the right: Beeswarm plots of SHAP-values for the explanatory variables for all models. Variables are sorted by their mean absolute SHAP value in descending order with the most important variables at the top. Each dot represents an observation in the study

Table 4 ROC AUC (95% CI) performance comparison of the four models applied to the training and test sets. 95% CI is computed with 2000 stratified bootstrap replicates

Model	Training	Test
LR	0.86 (0.85–0.88)	0.86 (0.83–0.89)[a]
RF	0.998 (0.997–1)	0.87 (0.85–0.9)[a]
NNET	0.87 (0.86–0.89)	0.87 (0.84–0.89)[a]
SVM	0.96 (0.95–0.97)	0.80 (0.76–0.84)[b]

[ab]Different letters in the same column indicate significant statistical differences ($p < 0.05$, Delong Test)

highest TPR (TPR = 0.90 while TPR = 0.87 and 0.83 for RF and LR). RF had the best area under the PR curve (PR AUC = 0.39), followed by LR (PR AUC = 0.38) and NNET (PR AUC = 0.37). Brier score did not show any difference among models. SVM had the worst performance in terms of all metrics.

When non-linear effects were incorporated into LR via restricted cubic splines, there was more overfitting, and the performance of LR was not increased compared to traditional LR (see supplementary material).

Figure 10 highlights that the best-performing models give similar predictions. LR, RF, and NNET predictions are strongly correlated (LR-RF, 0.84; LR- NNET, 0.92; RF- NNET, 0.92). On the other hand, the correlations of LR, RF, and NNET predictions with SVM ditto are weak. The predictions for the observed extreme events only follow the same pattern: high correlation between LR and RF, NNET, and low correlation with SVM. The distribution of predicted probabilities for non-extreme events coincides with the expected left-skewed pattern in all models. In contrast, the distribution of the expected probabilities for extreme event occurrences deviates from the ideal case of a highly right-skewed distribution.

5 Discussion

5.1 Physical interpretation and importance of the explanatory variables

We found that CAPE, TCW, some PCs of Z500, and the interaction term of CAPE and SATmax are significant variables for explaining the occurrence of extreme precipitation events in the Copenhagen area. As deduced theoretically, CAPE and TCW are the most important variables. Larger values of CAPE mean a larger amount of potential energy to be released by convection, and we expect this to have a positive effect on the probability of an event. Our analysis confirms this since the confidence interval of the odds ratio of CAPE is entirely above unity. Furthermore, also TCW has a confidence interval for the odds ratio, which is entirely above unity, which is as expected since we expect a larger TCW to mean a larger probability of extreme precipitation. The fact that these conclusions for CAPE and TCW are stable under

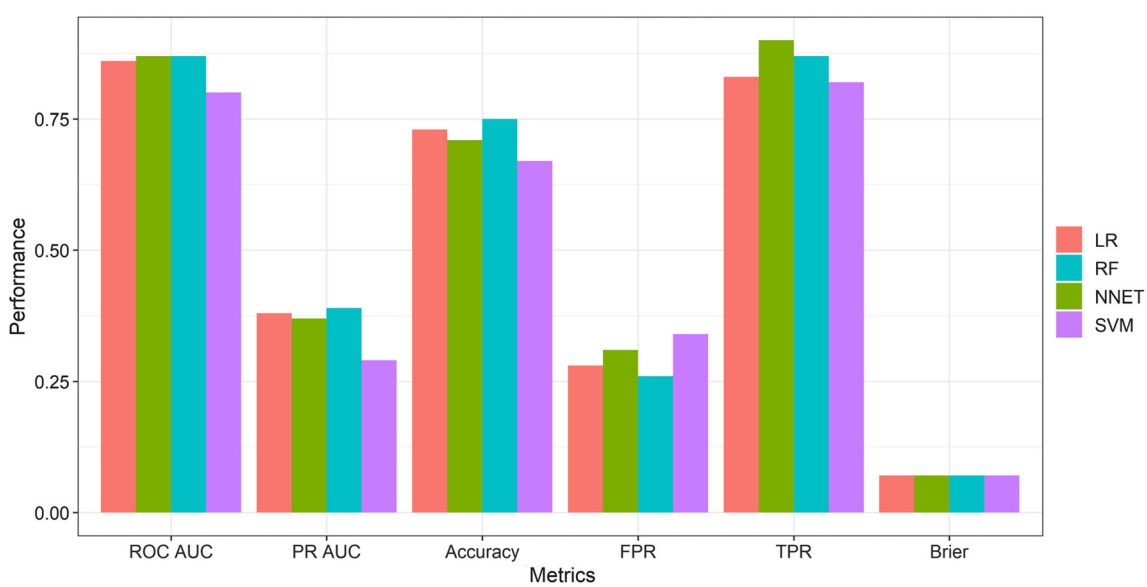


Fig. 9 Performance metrics of four models predicting the outcomes of extreme precipitation events

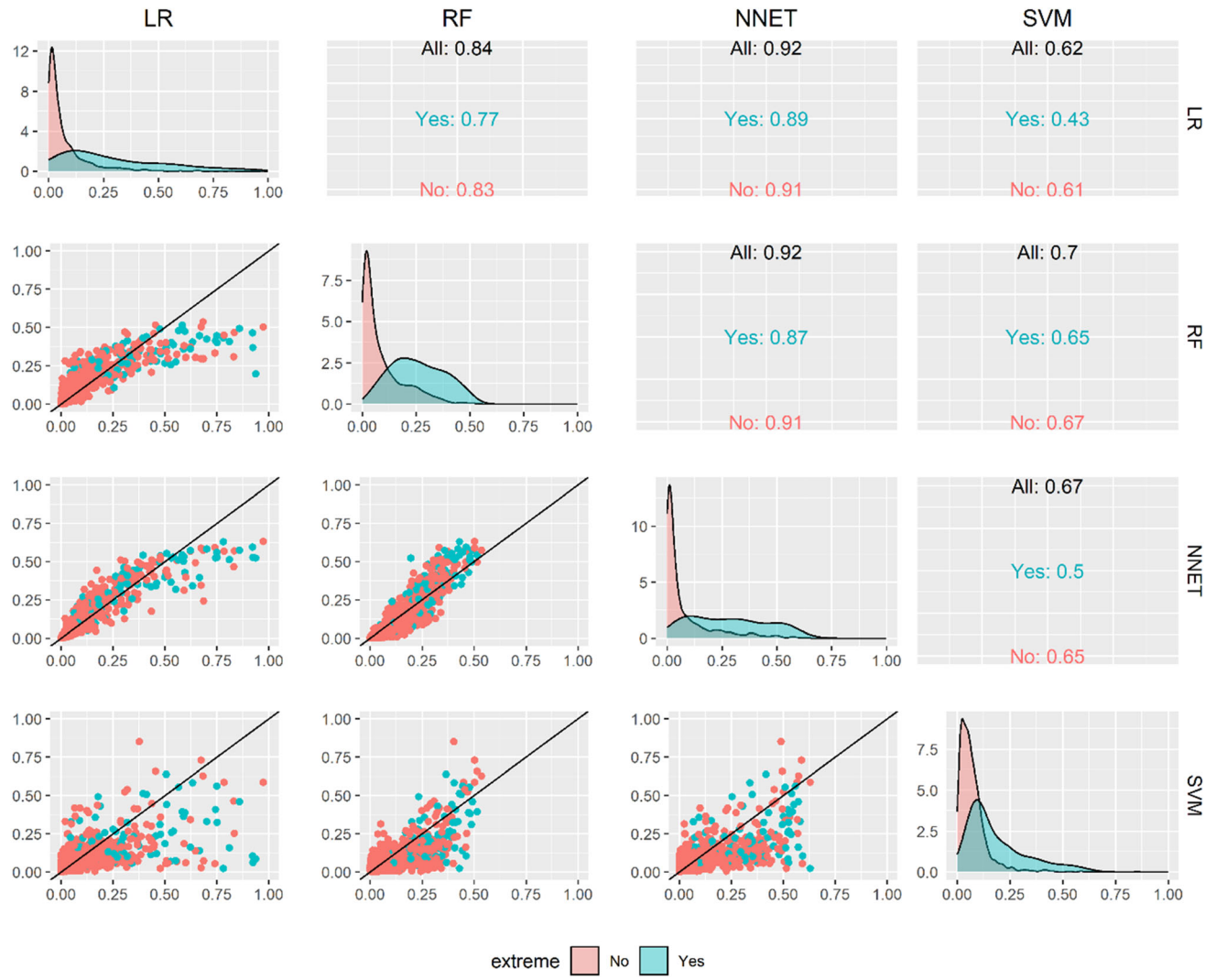


Fig. 10 Correlation coefficients for all data, extremes, non-extremes, and scatter plots of the probabilistic model predictions between the different models. In the diagonal are the density plots of probabilistic predictions of extremes/non-extremes for each model

subsampling and that these explanatory variables were found to be the most important variable by both established and more modern variable importance metrics for LR and ML models underlines the robustness of this conclusion. The rank of these explanatory variables may differ (e.g., CAPE is the first for NNET, RF, and second for LR, SVM) because each algorithm has a different approach to modeling the relationship between the explanatory variables and the response.

Furthermore, some of the Z500 PCs have odds ratios different from unity. A closer examination of these reveals that easterly and southeasterly flow favors extreme events, even on days where values of CAPE and TCW are low. We interpret this as already active convective systems being advected into the areas. This is, however, a less dominant mechanism since the PCs have much lower variable importance than CAPE and TCW.

We regarded SST as an explanatory variable of extreme precipitation events. However, after incorporating TCW, we found that SST had no significant impact. This is probably because SST variability is too small to influence the formation of extreme precipitation events, and that SST and TCW are correlated. On the other hand, SATmax improved the PR AUC slightly, but different models disagree on its importance.

The odds ratio for the interaction term of CAPE and SATmax is less than unity, which is surprising. We would have expected that the combination of large CAPE and high SATmax was in favor of releasing extreme events, but this conflicts with our results. The fact that the ML models also learn the same relationship consolidates this finding.

It is interesting to note that the most important explanatory variables; (CAPE, TCW, and Z500 PC's) all describe the state of the free atmosphere over the region

without regard to the detailed geography. This suggests that the models presented here and the conclusions about their comparison can be applied to other regions (with other fitted parameter values). Also, the physical arguments given earlier for choosing these as potential explanatory variables support this view. Given the limitations in the variable selection, it is possible to argue that expanding the set of variables used in the ML models could potentially yield different results in terms of variable importance. Including more variables may uncover additional interactions that were not captured by the original subset of variables.

The model with PC Z500, CAPE, SATmax & TCW as explanatory variables had high scores in all performance metrics except PR AUC. The reason for this is the high FPR values. Although we have a dense network of rain gauges in this case study, it is still possible to miss events that actually occurred. For the Copenhagen case, we would, e.g., miss any cloudburst that hits right off the coast in the sea. This fact could contribute to high FPR values. To validate this hypothesis, we conducted the analysis using data just from one gauge (the gauge with the least missing data). The results showed (see supplementary material) that the PR AUC is significantly lower even if the ROC AUC is comparable with the full network performance. These findings support our claim for higher FPR values attributed to missed events. Future work should therefore focus on augmenting the database by testing the sensitivity towards defining a region, e.g., by using fewer gauges and/or using bias-corrected weather radar data.

5.2 LR versus ML

We hypothesized that ML would be superior to traditional LR in predicting extreme events. Our hypothesis was partially supported in that LR performed worse than RF and NNET but better than SVM. Differences in the AUC between LR and the best-performing algorithms were not statistically significant, indicating only minor differences in overall fair explanatory performance across all models. Moreover, LR and NNET did not exhibit overfitting in contrast to RF and SVM. LR models might be more beneficial in this study than ML methods due to their transparency and interpretability. LR models also have a solid theoretical background, which allows for the use of well-defined statistical tests to assess the statistical significance of the explanatory variables.

Our findings may have implications for the design of future ML applications. Despite the simple model, data-driven ML methods yielded to a small performance benefit compared to the LR model. Finally, the difficulty of interpreting ML models is a barrier to their use in extreme event prediction. Future research could extract information from the black box to create interpretable but accurate statistical models. Each dataset is unique, and there is no 'free lunch' (Wolpert and Macready 1997), i.e., an algorithm performing well on a particular class of problems. Evaluating multiple algorithms against LR is critical to see if one outperforms the other; if performance is comparable, the simplest and most interpretable model should be employed.

5.3 Strengths and limitations

A systematic comparison between LR and several machine learning algorithms was conducted with a focus on their suitability in our setting of studying extreme precipitation events. Performance metrics specifically designed for unbalanced data were employed, and the ML models were optimized through a grid search approach. An independent dataset was utilized as the test dataset to enhance the validity of the findings, and variable importance metrics for all models were also employed to complement one another.

Extreme precipitation events are rare occurrences by nature, resulting in a limited amount of available data for training and evaluating models. The presented case area has an excellent climatological data set of 15 rain gauges within a relatively small geographical area and 42 years of recordings, yet the total number of observed extreme events that we are training models on is just 557. The lack of comprehensive, long-term observations is a general problem when studying the climatology of extreme rainfall, and relevant methods have to deal with this. The scarcity of extreme events restricts the number of variables that can reasonably be included in the analysis and the ability to employ very complex models. This is due to identifiability issues and that the risk of overfitting increases (Kuhn and Johnson 2013). In this study, we have thought it important to leverage domain expertise for selecting and engineering features that capture various aspects of extreme event behavior. Regarding variable selection, this study's approach may favor the LR model to some extent, as the process of selecting variables for ML models aligns with that of the LR model. This makes the model performance comparison straightforward and allows us to investigate how well LR and ML models are able to utilize information based on a mix of data and domain expertise. However, this may also be considered a limitation of our study since one of the most significant advantages of ML over LR is its ability to model complex, non-linear relationships between several explanatory variables and outcomes. The explanatory variables-outcome complexity should not be too low for ML to provide a meaningful advantage over LR, which is likely one the main reasons why the ML models are not able to outperform LR here. We hypothesize that this is likely to hold for other similar studies due to the lack of

comprehensive observational records. For ML methods, it has been stated as a rule of thumb that almost ten times as many events per variable are required to get consistent results compared to traditional statistical modeling (van der Ploeg et al. 2014). Future research could delve more into how far we can push the number of features in different types of ML models without running into lack of identifiability and over-fitting issues in studies of extreme precipitation events. It would also be interesting to see if ML models with more variables allow for physically interpretable insights that enhance our understanding of the climate system, such as drawing clear conclusions on variable importance. This introduces the challenge of deciding which variable selection method to employ for the ML models, potentially leading to variations in the selected variables and confounding a fair comparison across models.

Additionally, it is important to note that the neural network model used in this study is not particularly complex, as the cross-validation process limited the size of the hidden layer to only two. This could reduce the ability of the neural network to capture more complex relations in the data, but increasing the depth of the network also leads to lack of identifiability and over-fitting on small data sets.Furthermore, the LR depends on the fulfillment of specified assumptions (i.e., observation independence, no multicollinearity). Violating these may impact the quality of the analysis. The results of this study imply that any assumption violations with respect to LR do not significantly influence the performance quality because LR performed quite similarly to ML. Since non-linear effects were tested for LR via cubic splines without increasing the performance, their difference in performance is likely due to the lack of inclusion of variable interactions, which have to be included manually in LR by the user, while the ML models capture them automatically.

Lastly, CAPE, Z500, and TCW are simulated reanalysis outputs, which come with their own uncertainties and errors. One source of uncertainty is the used data assimilation method since different methods can yield to different results. Therefore the choice of method can impact the accuracy of the reanalysis data. Another source of uncertainty is the quality and consistency of the used observations. Older observations might not be as accurate as current ones because the observational methods have changed. This can impact the quality of the reanalysis data, particularly for earlier periods. Finally, the numerical models to produce the reanalysis data, including their spatial and temporal resolution, also have their own uncertainties.

6 Conclusion

In conclusion, the logistic regression framework was found to be an effective tool in modeling the occurrence of extreme precipitation events using meteorological drivers as explanatory variables. Considerable effort was put into improving model performance, generalization, and interpretation. The results showed that CAPE and TCW were the most important explanatory variables, while some of the PCs of Z500 and the interaction term of CAPE and SATmax were also significant. Our analysis confirmed the relationship between larger values of CAPE and TCW and a higher probability of extreme precipitation. The results also indicated that certain flow directions are favorable for extreme events. The results showed that SST had no significant impact, while SATmax improved the PR AUC slightly. The interaction term of CAPE and SATmax was found to be less than unity, which was surprising and requires further investigation. The model with PC Z500, CAPE, SATmax, and TCW as explanatory variables had high scores in most performance metrics, but high FPR values indicate a need to augment the precipitation data in future studies.

A classic LR performs similarly to more complex ML algorithms in a classification setting with four explanatory variables. This study demonstrates the value of comparing standard regression modeling to ML, mainly when a small number of well-understood, strong explanatory variables are used. Given the increasing availability of data technologies, ML may play a more significant role in prediction in the future. However, we still recommend caution in the optimism of using ML since its benefits depend on various criteria such as sample size, number of explanatory variables, and the complexity of the interactions of the explanatory variables. All of which are significant obstacles when working with rare extreme events and limited observational records.

In this study, a simple dataset gives reliable information on what circumstances lead to hourly precipitation extremes. Application to other locations to test transferability in both model structure and actual model remains to be tested, but early results are promising. This approach could be useful in data-sparse regions or for predicting the impacts of climate change where the physical understanding of convective rainfall is limited.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00477-023-02511-3.

Acknowledgements The observational precipitation dataset is a product of The Water Pollution Committee of The Society of Danish Engineers made freely available for research purposes. Access to data is governed by the Danish Meteorological Institute, and they should

be contacted for inquiries regarding data access. The ERA5 data used are available on the ECMWF MARS server.

Author contributions NA: Conceptualization, Writing—Original Draft, Investigation, Methodology, Data Curation, Visualization. HJDS: Conceptualization, Writing—Review & Editing, Methodology, Visualization, Validation. JWP: Conceptualization, Writing—Review & Editing, Methodology, Visualization, Validation. IBG: Conceptualization, Review, Methodology, Validation. TS: Conceptualization, Supervision, Writing—Review & Editing, Methodology, Visualization, Validation. KA-N: Conceptualization, Supervision, Writing—Review & Editing, Methodology, Project Management, Funding acquisition.

Funding Open access funding provided by Technical University of Denmark. Nafsika Antoniadou received funding from the Danish State through the National Centre for Climate Research (NCKF).

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Benedetti R (2010) Scoring rules for forecast verification. Mon Weather Rev 138:203–211. https://doi.org/10.1175/2009MWR2945.1

Bishop CM (2006) Pattern recognition and machine learning, Information science and statistics. Springer, New York

Boulesteix A-L, Schmid M (2014) Machine learning versus statistical modeling: machine learning versus statistical modeling. Biom J 56:588–593. https://doi.org/10.1002/bimj.201300226

Breiman L (2001a) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci. https://doi.org/10.1214/ss/1009213726

Breiman L (2001b) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Budach L, Feuerpfeil M, Ihde N, Nathansen A, Noack N, Patzlaff H, Naumann F, Harmouch H (2022) The effects of data quality on machine learning performance. https://doi.org/10.48550/ARXIV.2207.14529

Chan SC, Kendon EJ, Roberts N, Blenkinsop S, Fowler HJ (2018) Large-scale predictors for extreme hourly precipitation events in convection-permitting climate simulations. J Clim 31:2115–2131. https://doi.org/10.1175/JCLI-D-17-0404.1

Chen R-C, Dewi C, Huang S-W, Caraka RE (2020) Selecting critical features for data classification based on machine learning methods. J Big Data 7:52. https://doi.org/10.1186/s40537-020-00327-4

Coles S (2001) An introduction to statistical modeling of extreme values, Springer series in statistics. Springer, London. https://doi.org/10.1007/978-1-4471-3675-0

Cox DR, Snell EJ, Cox DR, Snell EJ (1999) Analysis of binary data, 2. ed., 1. CRC Press reprint. ed, Monographs on statistics and applied probability. Chapman & Hall [u.a.], Boca Raton

Cristianini N, Shawe-Taylor J (2000) An Introduction to support vector machines and other kernel-based learning methods, 1st edn. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511801389

Davenport FV, Diffenbaugh NS (2021) Using machine learning to analyze physical causes of climate change: a case study of U.S. midwest extreme precipitation. Geophys Res Lett. https://doi.org/10.1029/2021GL093787

Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on machine learning—ICML '06. Presented at the the 23rd international conference. ACM Press, Pittsburgh, pp 233–240. https://doi.org/10.1145/1143844.1143874

Dawson A (2016) eofs: a library for EOF analysis of meteorological, oceanographic, and climate data. JORS 4:14. https://doi.org/10.5334/jors.122

DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845

Deo RC, Nallamothu BK (2016) Learning about machine learning: the promise and pitfalls of big data and the electronic health record. Circ Cardiovasc Quality Outcomes 9:618–620. https://doi.org/10.1161/CIRCOUTCOMES.116.003308

Dittus AJ, Karoly DJ, Donat MG, Lewis SC, Alexander LV (2018) Understanding the role of sea surface temperature-forcing for variability in global temperature and precipitation extremes. Weather Clim Extremes 21:1–9. https://doi.org/10.1016/j.wace.2018.06.002

DMI (2015) Baltic sea-sea surface temperature reprocessed. https://doi.org/10.48670/MOI-00156

Fernández A (2018) Learning from imbalanced data sets. Springer, New York

Folland CK, Knight J, Linderholm HW, Fereday D, Ineson S, Hurrell JW (2009) The summer North Atlantic oscillation: past, present, and future. J Clim 22:1082–1103. https://doi.org/10.1175/2008JCLI2459.1

Gauthier J, Wu QV, Gooley TA (2020) Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. Bone Marrow Transplant 55:675–680. https://doi.org/10.1038/s41409-019-0679-x

Gevrey M, Dimopoulos I, Lek S (2003) Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol Model 160:249–264. https://doi.org/10.1016/S0304-3800(02)00257-0

Greenwell BM, Boehmke BC (2020) Variable importance plots—an introduction to the vip package. R J 12:343. https://doi.org/10.32614/RJ-2020-013

Gregersen IB, Madsen H, Rosbjerg D, Arnbjerg-Nielsen K (2013a) A spatial and nonstationary model for the frequency of extreme rainfall events: modeling the frequency of extreme events. Water Resour Res 49:127–136. https://doi.org/10.1029/2012WR012570

Gregersen IB, Sørup HJD, Madsen H, Rosbjerg D, Mikkelsen PS, Arnbjerg-Nielsen K (2013b) Assessing future climatic changes of rainfall extremes at small spatio-temporal scales. Clim Change 118:783–797. https://doi.org/10.1007/s10584-012-0669-0

Guth S, Sapsis TP (2019) Machine learning predictors of extreme events occurring in complex dynamical systems. Entropy 21:925. https://doi.org/10.3390/e21100925

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn, Springer series in statistics. Springer, New York

Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A, Soci C, Abdalla S, Abellan X, Balsamo G, Bechtold P, Biavati G, Bidlot J, Bonavita M, Chiara G, Dahlgren P, Dee D, Diamantakis M, Dragani R, Flemming J, Forbes R, Fuentes M, Geer A, Haimberger L, Healy S, Hogan RJ, Hólm E, Janisková M, Keeley S, Laloyaux P, Lopez P, Lupu C, Radnoti G, Rosnay P, Rozum I, Vamborg F, Villaume S, Thépaut J (2020) The ERA5 global reanalysis. Q J R Meteorol Soc 146:1999–2049. https://doi.org/10.1002/qj.3803

Hertig E, Jacobeit J (2013) A novel approach to statistical downscaling considering nonstationarities: application to daily precipitation in the Mediterranean area: Downscaling Under Nonstationarities. J Geophys Res Atmos 118:520–533. https://doi.org/10.1002/jgrd.50112

Hertig E, Seubert S, Paxian A, Vogt G, Paeth H, Jacobeit J (2014) Statistical modelling of extreme precipitation indices for the Mediterranean area under future climate change: statistical modelling of extreme precipitation. Int J Climatol 34:1132–1156. https://doi.org/10.1002/joc.3751

James G, Witten D, Hastie T, Tibshirani R (2021) An introduction to statistical learning: with applications in R, Second edition. Springer texts in statistics. Springer, New York. https://doi.org/10.1007/978-1-0716-1418-1

Jonkman SN (2005) Global perspectives on loss of human life caused by floods. Nat Hazards 34:151–175. https://doi.org/10.1007/s11069-004-8891-3

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on artificial intelligence—volume 2, IJCAI'95. Morgan Kaufmann Publishers Inc., San Francisco, pp 1137–1143

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Soft. https://doi.org/10.18637/jss.v028.i05

Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York

Lee J, Kim J, Lee J-H, Cho I-H, Lee J-W, Park K-H, Park J (2012) Feature selection for heavy rain prediction using genetic algorithms. In: The 6th international conference on soft computing and intelligent systems, and the 13th international symposium on advanced intelligence systems. Presented at the 2012 joint 6th international conference on soft computing and intelligent systems (SCIS) and 13th international symposium on advanced intelligent systems (ISIS), IEEE, Kobe, Japan, pp 830–833. https://doi.org/10.1109/SCIS-ISIS.2012.6505383

Lepore C, Veneziano D, Molini A (2015) Temperature and CAPE dependence of rainfall extremes in the eastern United States. Geophys Res Lett 42:74–83. https://doi.org/10.1002/2014GL062247

Li J, Wang B (2018) Predictability of summer extreme precipitation days over eastern China. Clim Dyn 51:4543–4554. https://doi.org/10.1007/s00382-017-3848-x

Lindsey JK (2000) Applying generalized linear models, Corr. 3. printing. ed, Springer texts in statistics. Springer, New York

Liu JNK, Li BNL, Dillon TS (2001) An improved naive Bayesian classifier technique coupled with a novel input solution method [rainfall prediction]. IEEE Trans Syst Man Cybern C 31:249–256. https://doi.org/10.1109/5326.941848

Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. https://doi.org/10.48550/ARXIV.1705.07874

Maidens A, Knight JR, Scaife AA (2021) Tropical and stratospheric influences on winter atmospheric circulation patterns in the North Atlantic sector. Environ Res Lett 16:024035. https://doi.org/10.1088/1748-9326/abd8aa

Mason SJ, Graham NE (2002) Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Q J R Meteorol Soc 128:2145–2166. https://doi.org/10.1256/003590002320603584

Mastrantonas N, Herrera-Lormendez P, Magnusson L, Pappenberger F, Matschullat J (2021) Extreme precipitation events in the Mediterranean: Spatiotemporal characteristics and connection to large-scale atmospheric flow patterns. Int J Climatol 41:2710–2728. https://doi.org/10.1002/joc.6985

Mayer M, Watson D (2023) kernelshap: Kernel SHAP. https://CRAN.R-project.org/package=kernelshap

Merino A, Sánchez JL, Fernández-González S, García-Ortega E, Marcos JL, Berthet C, Dessens J (2019) Hailfalls in southwest Europe: EOF analysis for identifying synoptic pattern and their trends. Atmos Res 215:42–56. https://doi.org/10.1016/j.atmosres.2018.08.006

Meyer H, Kühnlein M, Appelhans T, Nauss T (2016) Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. Atmos Res 169:424–433. https://doi.org/10.1016/j.atmosres.2015.09.021

Mitchell TM (1997) Machine learning, McGraw-Hill series in computer science. McGraw-Hill, New York

Moon S-H, Kim Y-H, Lee YH, Moon B-R (2019) Application of machine learning to an early warning system for very short-term heavy rainfall. J Hydrol 568:1042–1054. https://doi.org/10.1016/j.jhydrol.2018.11.060

Nembrini S, König IR, Wright MN (2018) The revival of the Gini importance? Bioinformatics 34:3711–3718. https://doi.org/10.1093/bioinformatics/bty373

O'brien RM (2007) A caution regarding rules of thumb for variance inflation factors. Qual Quant 41:673–690. https://doi.org/10.1007/s11135-006-9018-6

Oehlert GW (1992) A note on the Delta method. Am Stat 46:27. https://doi.org/10.2307/2684406

Pawitan Y (2014) In all likelihood: statistical modelling and inference using likelihood. OUP, Oxford

Pedersen AN, Mikkelsen PS, Arnbjerg-Nielsen K (2012) Climate change-induced impacts on urban flood risk influenced by concurrent hazards: climate change-induced impacts on urban flood risk. J Flood Risk Manag 5:203–214. https://doi.org/10.1111/j.1753-318X.2012.01139.x

Ramezani Ziarani M, Bookhagen B, Schmidt T, Wickert J, de la Torre A, Hierro R (2019) Using convective available potential energy (CAPE) and dew-point temperature to characterize rainfall-extreme events in the South-Central Andes. Atmosphere 10:379. https://doi.org/10.3390/atmos10070379

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform 12:77. https://doi.org/10.1186/1471-2105-12-77

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:e0118432. https://doi.org/10.1371/journal.pone.0118432

Scaife AA, Folland CK, Alexander LV, Moberg A, Knight JR (2008) European climate extremes and the North Atlantic oscillation. J Clim 21:72–83. https://doi.org/10.1175/2007JCLI1631.1

Shi X (2020) Enabling smart dynamical downscaling of extreme precipitation events with machine learning. Geophys Res Lett. https://doi.org/10.1029/2020GL090309

Solantie R, Frisk K, Croitoru A-E (2006) Major summer cloudbursts in Finland: synoptic origins and impact. Wea 61:159–163. https://doi.org/10.1256/wea.274.04

Storch HV, Zwiers FW (1984) Statistical analysis in climate research, 1st edn. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511612336

Strobl C, Boulesteix A-L, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform 8:25. https://doi.org/10.1186/1471-2105-8-25

Sun B, Wang H (2018) Interannual variation of the spring and summer precipitation over the three river source region in china and the associated regimes. J Clim 31:7441–7457. https://doi.org/10.1175/JCLI-D-17-0680.1

Thomassen ED, Thorndahl SL, Andersen CB, Gregersen IB, Arnbjerg-Nielsen K, Sørup HJD (2022) Comparing spatial metrics of extreme precipitation between data from rain gauges, weather radar and high-resolution climate model re-analyses. J Hydrol 610:127915. https://doi.org/10.1016/j.jhydrol.2022.127915

van der Ploeg T, Austin PC, Steyerberg EW (2014) Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 14:137. https://doi.org/10.1186/1471-2288-14-137

Vicente-Serrano SM, Beguería S, López-Moreno JI, El Kenawy AM, Angulo-Martínez M (2009) Daily atmospheric circulation events and extreme precipitation risk in northeast Spain: role of the North Atlantic Oscillation, the Western Mediterranean Oscillation, and the Mediterranean Oscillation. J Geophys Res 114:D08106. https://doi.org/10.1029/2008JD011492

Wei W, Yan Z, Jones PD (2020) A decision-tree approach to seasonal prediction of extreme precipitation in eastern China. Int J Climatol 40:255–272. https://doi.org/10.1002/joc.6207

Wilks DS (2011) Statistical methods in the atmospheric sciences, 3rd edn, International geophysics series. Academic Press, Oxford

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Computat 1:67–82. https://doi.org/10.1109/4235.585893

Wulff CO, Greatbatch RJ, Domeisen DIV, Gollan G, Hansen F (2017) Tropical forcing of the summer east Atlantic pattern. Geophys Res Lett. https://doi.org/10.1002/2017GL075493

Xoplaki E, González-Rouco JF, Luterbacher J, Wanner H (2004) Wet season Mediterranean precipitation variability: influence of large-scale dynamics and trends. Clim Dyn 23:63–78. https://doi.org/10.1007/s00382-004-0422-0

Liu X-Y, Jianxin Wu, Zhou Z-H (2009) Exploratory undersampling for class-imbalance learning. IEEE Trans Syst, Man Cybern B 39:539–550. https://doi.org/10.1109/TSMCB.2008.2007853

Yang Y, Huang F, Wang H (2013) Dominant modes of geopotential height in the northern hemisphere in summer on interdecadal timescales. Chin J Ocean Limnol 31:1120–1128. https://doi.org/10.1007/s00343-013-2229-5

Ziersen J, Clauson-Kaas J, Rasmussen J (2017) The role of Greater Copenhagen Utility in implementing the city's Cloudburst Management Plan. Water Pract Technol 12:338–343. https://doi.org/10.2166/wpt.2017.039

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.