Authors' reply on Manuscript Predicting power outages caused by extratropical storms

Corresponding author: Roope Tervo, roope.tervo@fmi.fi
Oct 30th, 2020

We thank all referees for taking the time to read our paper and for giving us insightful, constructive, and extremely valuable comments and improvement suggestions. We have addressed all the comments as accurately and precisely as possible and made the improvements in the manuscript.

In the following, we respond to each referee's comments item-by-item. The referee's comments are indented and with italic typesetting. The authors' comments are with normal typesetting. Direct quotes from the manuscripts are marked with double-quotes.

The revised manuscript is also attached along with a "track-changes version" with all changes highlighted. The page and line numbers in the responses refer to the revised manuscript (not the track-changes version).

Table of contents:

Referee 1

Anonymous referee

Responds to the general remarks

General remarks

The article investigates windstorm impacts on the power grid in Finland. The authors present a methodology to identify storm objects as polygons and combine them with meteorological and non-meteorological data to predict power outages. They use ERA5 reanalysis data, a national forest inventory and a dataset with information about time and location of power outages in Finland. Storm objects are identified using a fixed wind speed threshold of 15 m/s are tracked in time and space. A large set of meteorological and non-meteorological parameters is gathered for each storm object. From these parameters the most relevant are selected and five different methods are used to classify the storm objects with respect to the damage they caused to the power grid using three damage classes. It is tested how well the different methods are able to predict the class of a storm object using cross-validation. Finally, the best performing classification method is applied to three test cases of severe storms.

In general, the article addresses the very interesting and relevant topic of predicting the impacts of extreme weather events. The authors use state-of-the-art data and methodology. However, there are some issues in the manuscript and there are some parts that need more detailed explanation and discussion. These issues should be addressed before the manuscript is accepted.

The authors use sophisticated methods for classification of storm objects with a large set of parameters. What is missing in the study is ***an analysis of the relevance of the individual parameters for the classification task.*** *It remains unclear* ***which of the parameters play an important role****. It might be, for example, that it is mainly the size of the storm object or the number of transformers under the object that is relevant for the damage, while the standard deviation of wind direction plays a minor role. It would be beneficial* ***to include an analysis of the importance of the parameters, at least for the best performing method, to add more scientific insight to the rather technical aspects of classification task.***

> We conducted a permutation feature importance analysis using the Gaussian processes (GP) model and a randomly selected test set of the national dataset. The same model and data are used to produce the case examples.
>
> The manuscript is appended with the following chapters (page 17 in the updated manuscript):

“The relevance of the individual predictive features can be explored by using the permutation test, as done by Breiman (2001). First, the baseline score of the fitted model is calculated using the test set. Then each feature is randomly permuted, and the difference in the scoring function is calculated. The random permutation is repeated 30 times for each parameter, and the average of the results is used. The procedure offers information on how important the feature is to obtain good results. It should be mentioned that highly correlated features may get low importance as other features work as a proxy to the permuted feature. However, using completely independent features is not possible in weather data since weather parameters are often dependent on each other, and eliminating even the most apparent pairs from the used features impaired the results in our experiments.

We used the macro average of F1 defined in Equation 8 as a scoring function and the randomly selected test set from the national data. The relevance is shown in Figure 7. Most features show at least little relevance for the results. The first twelve features are significantly more relevant than the rest. The most important features contain at least one representative of all meteorological parameters used in training. In other words, all employed meteorological parameters are important for the prediction, while different aggregations are contributing to the "fine-tuning" of the model.

As Figure 7 shows, the most significant parameter regarding our model performance is the average wind speed. Numerous studies support our result of wind being the most important damaging factor (Virot et al., 2016; Valta et al., 2019; Jokinen et al., 2015). They are, however, highlighting the importance of maximum wind gusts instead of the average wind. Surprisingly, in our analysis, the wind gust speed does not belong to the most critical parameters. Instead, maximum mixed layer height, related to the wind gustiness, contributes crucially to the model performance. The dependencies between predictive features might be one reason for some parameters to have a lower rank in the results.

The stand mean diameter and height are the most important features regarding the forest parameters, which corresponds to our expectations. Previous studies also state these features to influence the wind damage in forests (Pellikka and Järvenpää, 2003) and hence indirectly electricity grids. As Pellikka and Järvenpää (2003) and Suvanto et al. (2016) discuss, also the age of the forest has an impact on storm damages. However, in the feature importance test, forest age does not seem to contribute significantly to the prediction outcome.

The most important object feature is the size of the object. Object movement speed and direction did not contribute strongly to the results. However, previous studies indicate that besides the size of the impacted area, the duration of strong winds – i.e., the movement speed of the system – influences also the amount of damage (Lamb and Knud, 1991).”

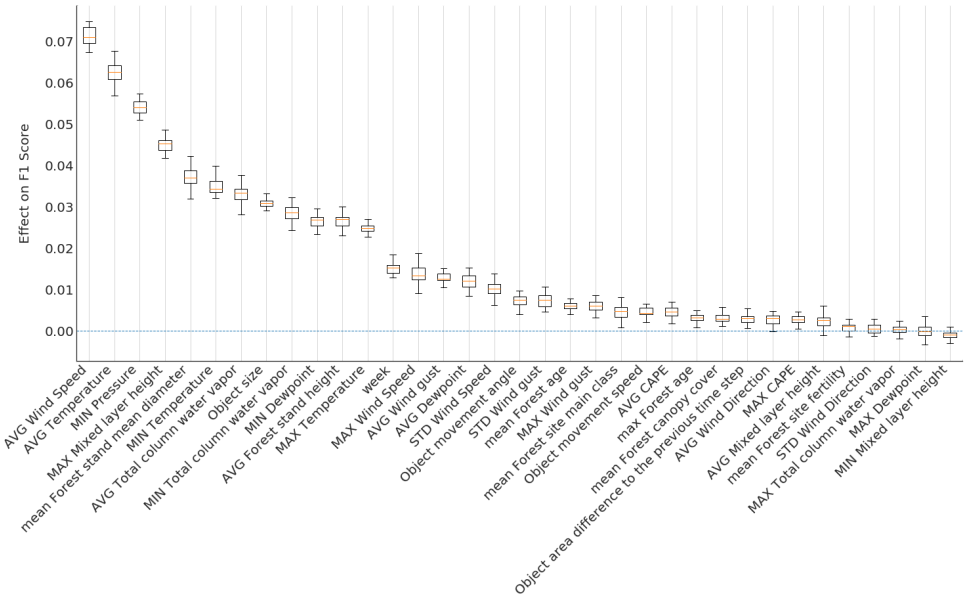


Figure 7. Permutation feature importance using the GP classification method trained with the randomly selected national dataset. The higher effect on the F1 score is (y-axis), the bigger is the significance.

The authors should discuss what is ***the benefit of using storm objects, rather than directly relating wind speeds and other parameters to power outages in a certain area, for example in a grid-based approach.*** *Following the approach in the manuscript, one is able to assign a damage class to the whole area of the storm object.* ***However, this does not provide any information about the specific location of the outage. I would suggest to discuss in more detail what could be the use of such large-scale damage information for an energy provider (see also my specific comment further below).***

Using storm-objects instead of fitting the models with gridded data is a fundamental design choice of the work. Its benefits and downsides will definitely be an interesting subject to cover. We added the following discussion into the manuscript (page 22, 408-423 in the updated manuscript):

“The presented object-based approach has both advantages and disadvantages. Extracting storm objects in advance prepro- cesses the data for machine-learning techniques, such as RFC, which do not perform feature learning. It enables machine-learning methods to focus only on the relevant parts of the data. Methods not containing feature learning, such as RFC and logistic regression, have been found to outperform neural networks for forest (Hart et al., 2019) and weather data (Tervo et al., 2019). It also leads to significantly faster training times. Processing objects instead of the grid makes it also easier to track and use object attributes such as age, speed, and movement. Moreover, objects are easy to visualize, and user interfaces may be enriched with related actions such as tracking and alarms.

On the other hand, storm objects use only aggregated attributes, which may decrease the classification accuracy when predictive features vary significantly under the storm object area. Several machine-learning methods, i.e., deep neural

networks, could be trained to employ those local features to gain better accuracy. Such methods could also utilize three-dimensional data. Extracting storm objects requires a fixed threshold of wind gust and pressure, which may vary depending on the characteristics of geospatial locations. Nevertheless, the previous studies indicate the critical threshold to wind gust speed to be the same for almost entire geospatial domain of this work (Gardiner et al., 2013). Moreover, the correct threshold may vary depending on the data source. When extending the geospatial domain or changing the data source, this might become a more serious issue, and different thresholds might be needed. One possibility to determine the optimal threshold might be to use specific quantiles of the parameter values, but this would need further investigation."

In many figures the labels are hardly readable.

We went carefully through all figures and enlarged the labels.

The manuscript needs to be checked for English language.

We carefully checked the language and made corrections to the manuscript.

Respond to the specific remarks

Page 3, line 81: What is the spatial resolution of the forest inventory?

This information has now been added to the manuscript on page 4, lines 97-100.

"The original geospatial resolution of the data is 16 meters, which has been reduced to approximately 1.6 km resolution to speed up the processing. Taking into account the size of extratropical cyclones (diameter 1000 km) and the wide areas where wind damages typically occur e.g. near to the cold front, we consider a resolution of 1.6 km being sufficiently high for modeling wind storm damages.."

Page 3, line 84-88: It could be useful to introduce Figure 1 already here in the data section. This would be helpful for the reader to understand the extraction of storm object feature in section 3.2. You should also go into more detail about the spatial accuracy of the local and national data set.

We have improved Figure 1 and moved it in the data section and made it more easily understandable and to have it in a more logical place. Firstly, we separated Figures 1a and 1b from 1c and 1d and improved the figures (pages 5 and page 9 in the updated manuscript). We have also added more information about the structure of the local and national dataset and on the spatial accuracy to the data section (page 4, lines 102-115):

“Power outage data are obtained from two complementary sources. The national dataset is acquired from the Finnish Energy (2010-2018) who aggregates the data from power distribution companies in Finland. The national data are provided only for research purposes and for areas containing a minimum of six grid companies; this is, for example, to ensure energy users’ anonymity. Therefore, the national dataset does not include exact locations of the faults. We have also obtained some parts of the data with better spatial accuracy from two individual power distribution companies. In this paper, we refer to this data as the local dataset. In the local dataset, the fault locations are reported in relation to transformers, i.e. the spatial resolution of the outages ranges from a few meters to kilometers.

Figure 1 illustrates the geographical coverage of the power outage data. The local dataset contains all outages from 2010 to 2018 in the northern area (Loiste) and outages related to major storms in the southern area (JSE), shown in Figure 1a. The national dataset contains all outages in Finland from 2010 to 2018 divided into five regions, shown in Figure 1b. The national dataset contains in total 6 140 434 outages with relatively low geographical accuracy. On the other hand, the local dataset repesents a substantially smaller geographical area with a good geographical accuracy but contains only 22 028 outages in total. We train our classification models, described in more detail in Chapter 3.4, with both datasets to evaluate their performance for different types of data.”

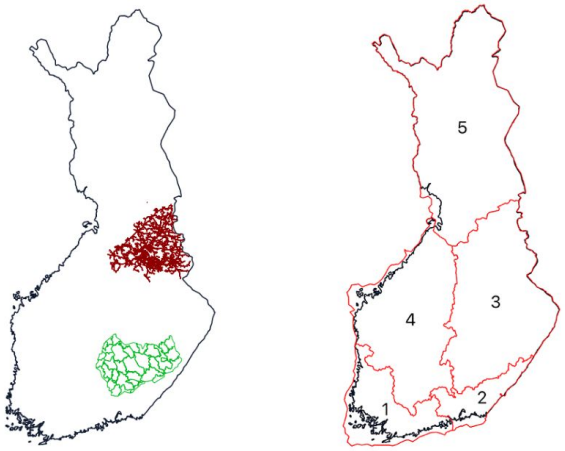


Figure 1. (a) Geographical coverage of the outage data (local dataset). The red lines represent the power grid of Loiste (northern grid company) and the green lines the operative areas of JSE (southern grid company). Outages of the local dataset are collected from both of these areas. (b) Regions in the national outage dataset. Outages are gathered from the whole Finland and aggregated to the regions shown in the figure.

Page 4, line 97: Can the storm polygons have "holes", if within the area of a polygon areas with winds below 15 m/s exist?

The contouring algorithm is capable of finding interior rings of the polygons. The used wind gust fields did not, however, contain any such cases. Thus one storm object represents a solid area (polygon).

This information has been added to the updated manuscript on page 5, lines 125-126.

page 4, line 103: Here you mention pressure objects for the first time. Are they defined by the 1000 hPa threshold? Please describe in more detail. Also, when you use the word "object" on its own, it is not clear if you refer to a "storm object" or "pressure object". Therefore you should only use "storm object" and "pressure object". Later you also use the term "wind object".

We clarified these paragraphs on page 5, lines 123-137, and revised the use of the word “object”. In particular, we describe the object identification and tracking method following:

“Storm objects are identified by finding contour lines of 10-meter wind gust fields using 15 m s−1 thresholds from the ERA5 surface level grid with a time step of 1 hour. The contouring algorithm is capable of finding interior rings of the polygons. The used wind gust fields did not, however, contain any such cases. Thus one storm object represents a solid area (polygon) where the hourly maximum wind gust exceeds 15 m s−1 during one particular hour. The threshold of 15 m s−1 is selected as different sources indicate Finland being vulnerable for windstorms and rather moderate winds (from 15 m s−1) causing damages to forests (Valta et al., 2019; Gardiner et al., 2013). Valta et al. (2019) developed a method to estimate the windstorm impacts on forests by combining the recorded forest damages from the nine most intense storms and their observed maximum inland wind gusts. According to the formula developed in the study, the inland wind gusts of 15 m s−1 alone result in forest damages of 1800 m^3. We also identify pressure objects by finding contour lines using a 1000 hPa threshold to connect potentially distant storm objects around the low-pressure center to the same storm event.

After identification, storm objects are tracked by connecting them with each other. Each storm object is first connected to nearby pressure objects from the current and preceding time steps. If pressure objects do not exist within the distance threshold, the object is connected to nearby storm objects from the current and preceding time steps. The Algorithm enables assigning each storm object to an overall event (low pressure system) and tracking the objects’ movement. Algorithm 1 shows the details of the process.”

page5, algorithm1: What is the"previous pressure object"? Is it previous in time? Or is there another for-loop that cycles through the pressure objects, which is not mentioned in the algorithm? What is"other object"? You mention "object", without specifying if it is a storm or pressure object. Please revise the algorithm, so that it is easy to understand for the reader.

The algorithm description has been updated to be more explicit. The readability may have been affected a little bit, but we believe this is a better and more precise way to

describe the tracking algorithm. Meritoriously notified questions about previous objects and object types are addressed as well.

The updated algorithm is listed below and updated to the manuscript.

Algorithm 1 Storm tracking

Input

Wind and pressure objects S_o arranged by time

$pressure\ distance\ threshold$

$wind\ distance\ threshold$

$speed\ threshold$

$time\ step$

Output

Connected wind and pressure objects with storm ID

for all wind and pressure object $O_{w|p} \in S_o$ **do**

$current\ time \leftarrow$ time of the object $O_{w|p}$

$previous\ time \leftarrow current\ time - time\ step$

Current time pressure objects $S_p^c \leftarrow$ pressure objects having centroid within $pressure\ distance\ threshold$ from object $O_{w|p}$ centroid and time stamp $current\ time$

Previous time pressure objects $S_p^p \leftarrow$ pressure objects having centroid within $speed\ threshold$ from object $O_{w|p}$ centroid and time stamp $previous\ time$

Current time wind objects $S_w^c \leftarrow$ wind objects having centroid within $wind\ distance\ threshold$ from object $O_{w|p}$ centroid and time stamp $current\ time$

Previous time wind objects $S_w^p \leftarrow$ wind objects having centroid within $speed\ threshold$ from object $O_{w|p}$ centroid and time stamp $previous\ time$

if pressure object $O_p^c \in S_p^c$ exists with ID **then**

Use pressure object $O_p^c\ ID$

else if pressure object $O_p^p \in S_p^p$ exists with ID **then**

Use previous time pressure object $O_p^p\ ID$

else if wind object $O_w^c \in S_w^c$ exists with ID **then**

Use wind object $O_w^c\ ID$

else if wind object $O_w^p \in S_w^p$ exists with ID **then**

Use previous time wind object $O_w^p\ ID$

else if wind or pressure object $O_{w|p}^p \in S_w^p \cup S_p^p$ exists without ID **then**

Give new ID to the previous object $O_{w|p}^p$ and current object $O_{w|p}$

else

Leave object $O_{w|p}^p$ without ID

end if

end for

page5, line123-128: From your description it is not clear how you selected the relevant parameters. You write about a fitted Gaussian distribution. How do you fit it, to which data and with which purpose? What is class one and two? What is the criterion for selecting the 35 relevant parameters?

We clarified the description as follows in the updated manuscript on pages 6-8, lines 156-173:

“We selected the 35 parameters based on two main criteria: First, we prepared a list of potential parameters detected in related 155 studies, e.g. Suvanto et al. (2016); Peltola et al. (1999); Valta et al. (2019), or identified through the empirical experience

of duty forecasters (Weather and Safety Center of Finnish Meteorological Institute - Duty forecasters, 05/2020). Second, we selected the relevant parameters, which were available to us or accessible with a reasonable effort. However, some possibly essential parameters, like soil temperature from ERA5 reanalysis, were left out because of the slow downloading process.

After the preliminary selection of the parameters, we conducted dozens of light experiments using different combinations of parameters and models to find the best possible setup. To this end, we fitted the Gaussian distribution to each parameter using at first all samples, then samples with few outages, and finally with many outages (classes 1 and 2 specified in Section 3.3). While many other distributions are known to suit better in modeling particular parameters, such as Gamma in precipitation, Weibull in wind speed, and Lognormal in cloud properties (Wilks, 2011), the Gaussian distribution is a sufficient simplification to help in selecting relevant parameters. We visually inspected the differences between fitted Gaussian distributions to deduce 165 the potential relevance of the parameter. Supposedly the distribution of one parameter is different for all samples and samples with many outages, and the classification method may exploit the parameter to predict the damage potential of the storm object. The distributions of some selected parameters are shown in Appendix A. In total, 35 parameters, shown as bold faced in Table 1, were chosen for the final classification."

page 7, line 130-131: At this point it is not clear how you define the three classes. To make it easier for the reader, I would suggest to spend some words on how the classes are defined here, or to move this part to page 8, line 155, where the classes are actually introduced.

We restructured the text to introduce classes on page 10, line 197 (originally on page 8, line 155), as you suggested.

page 7, line 136-138: You write "the local dataset contains 24,542 storm objects". Would it be more precise to say that "24,542 storm objects are related to outages in the local outage dataset"? It would be very informative to know how many outages are in the dataset in total and how many of them are NOT related to a storm object. Maybe you can add that information here.

Using only storm objects related to outages would result in overestimating predictions as the classification model would not see any "harmless" class 0 samples in the training process and assume every sample to cause damage. Thus, we also consider storm objects which are not related to any outage.

The local dataset contains 24 542 storm objects and 5 837 outages connected to 2 363 storm objects. Thus 22 179 storm objects in the local dataset have not caused any outages. The local power outage data contains 16 191 outages, which can not be connected to any storm object. The national dataset contains 142 873 storm objects and 5 965 324 outages connected to 33 796 storm objects. 109 077 storm

objects are not connected to any outages, and 175 110 outages can not be connected to any storm object.

We added this information to the manuscript on page 9, lines 170-175.

page 7, figure 1a: Can you explain why the network topologies look so different in the northern and southern area? In the north it looks like branches that end some where, in the south it rather looks like district boundaries. Figures 1 c and d: What is shown here in red color? Number of outages per area? Please add a legend. I would recommend to plot the grid topology with a darker color on top of the shading to increase its visibility.

The differences between the network topologies are simply explained by the data we have received from the two individual companies. From the northern company (Loiste), we received a shapefile of their grid. The southern company (JSE) provided their operational areas instead of the grid topology. Therefore, these two topologies look so different, even though in reality also JSE's grid looks similar compared to Loiste.

We have now separated Figures 1a and 1b from 1c and 1d and improved the figures based on the suggestions (Pages 5 and 9 in the updated manuscript). See also the reply to the second comment about the spatial accuracy of datasets.

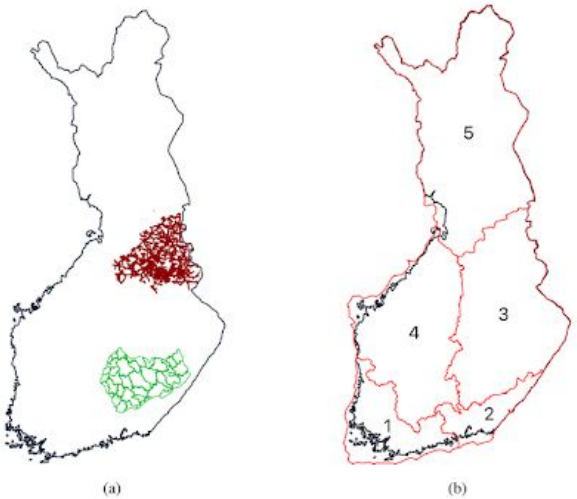


Figure 1. (a) Geographical coverage of the outage data (local dataset). The red lines represent the power grid of Loiste (northern grid company) district and the green lines the operative areas of JSE (southern grid company). Outages of the local dataset are collected from both of these areas. (b) Regions in the national outage dataset. Outages are gathered from the whole Finland but aggregated to the regions shown in the figure.

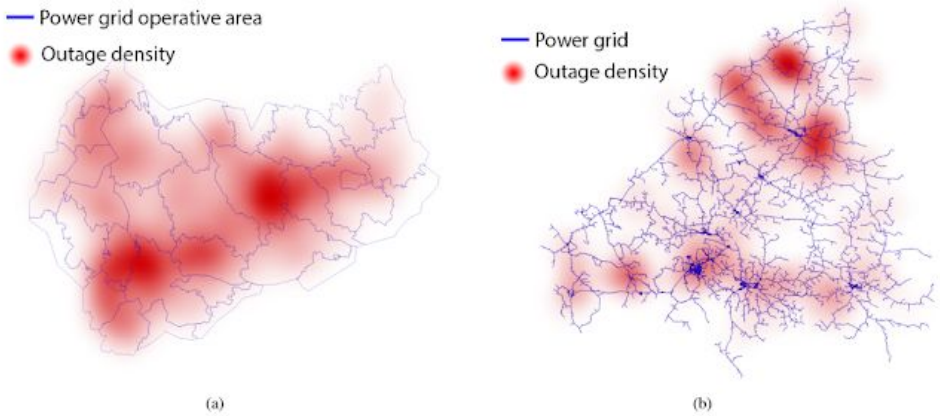


Figure 2. (a) Spatial distribution of the outages in the JSE network (southern area), data gathered between 2010 and 2018. (b) Spatial distribution of the outages in the Loiste network (northern area), data gathered between 2010 and 2018.

page8, line153-154: Please explain in more detail what is shown in figure 4. Does one dot represent the outages and affected customers related to a specific storm object? Is the line a linear regression?

One dot indeed represents the outages, but it may not be related to any specific storm. The line is trendline (linear regression). We also added a legend to the figure and extended the description in the manuscript on page 10, lines 192-196 following:

“Figure 5 renders how many customers are typically affected by one outage. The figure contains all outages in both datasets, whether they are related to a storm or not. In the local dataset, usually 20-30 customers lose electricity in one outage. In the national dataset, only six customers usually lose electricity in one outage. We assume that this roots to different network topologies between the areas. Notably, in some rare cases, a much higher number of customers are affected. Based on our random inspections, these cases occur typically in urban areas and are rare because the power network is mainly underground in these areas.”

The original manuscript also contained an error. The original manuscript stated that 200-300 would be typically affected, which is wrong. One outage usually affects from 6 to 30 outages depending on the dataset. We corrected this.

page 10, table 2: The caption say "Classes for local dataset", but shown are also classes for the national dataset.

We compliment, and corrected this on page 12, table 2.

page 10, line 153-154: Is "model" the correct term here? Isn’t it rather "classification algorithm"?

We assume that this refers to page 10, lines 163-164. The “model” is normally used in this context in machine-learning literature. We see the word “algorithm” to refer

more to heuristic algorithms instead of models that are fitted to the data. Another option would also be “method”, but it may be confused with the overall method, including storm identification and tracking.

We see that the word “model” is the best term in this context.

page 11, equations 1, 2, 3: If you use equations, you need to define the individual variables. Also, the equations are not easily understood without further explanation.

The definitions of the variables are fundamental for equations, and we added them to the manuscript. They should help to understand the equations. We also added references for all kernels used in this work. As the used kernels are widely used standard kernels, we prefer to omit a more detailed explanation to keep the text concise and readable.

page 14, section 4.1: As far as I can see it is not mentioned in the text which classification algorithm was used for the case examples.

Gaussian processes (GP) was used in case examples. Thus, we also analyzed feature importances using GP.

This information is added to the manuscript on page 19, line 349. We also changed a conclusion slightly on page 22, lines 403 to form:

“Both Gaussian Processes and Support Vector Classifiers provided good results. [...]”

The original statement in the conclusion honoured only SVC, which is inconsistent with results. SVC and GP provided almost similar performance.

page 15, figure 5: The figures should be as self-explanatory as possible. Please explain in the caption what the numbers represent.

We added the following information to the manuscript on page 18:

“Each cell of the confusion matrices represents a share of predictions having a corresponding combination of predicted and true class. For example, the middle right cell tells the share of samples belonging to class 1 but predicted to have class 2.”

page 16, line 305: The term "cell" is usually used for convective thunderstorms, but not for large-scale winter storms. I would suggest to simply use the word "storm".

Good point, this has been changed to “storm object”.

page 17, line 307: The authors state that "the model is able to provide a more specific and geospatially accurate prediction of caused damage to the power grid than for example

weather warning." I do not think that this statement is true. If I understand the model correctly, it assigns a damage class to the whole area of a storm object. This area can be quite large, as figure 6a and 6b show. Furthermore, the model provides no geospatial information about where inside this area the damages are expected. I suppose that weather warnings are available for Finland at a much higher spatial resolution. Additionally, weather warnings are released in advance of an event. In this manuscript the authors do not take into account forecast uncertainty. Therefore, a comparison to weather warnings difficult.

We acknowledge that the comparison with weather warnings can be challenging. As the referee mentions, the model's ability to provide more specific and geospatially accurate information than weather warnings is not a straightforward issue. We mention the geospatial accuracy because, in some cases, the storm object areas are not as big as in 6a and 6b (8a and b in updated manuscript), which are two examples of extremely strong storms. This was the case, for instance, with the extratropical storm, Pauliina where the yellow level of wind warnings was issued to wide areas in central and southern Finland and orange level of wind warnings to the south (see attached figure). This broad wind warning likely leads to all power companies in southern and central Finland being alert and possibly overpreparing for the event. Another important aspect of this work compared to weather warnings is an analysis of inflicted power outages, which can give an insight to power grid operators and duty forecasters about the impacts of forecasted warnings.

Nevertheless, because of the problematic task to indeed take into account the uncertainty of the forecast, we decided to modify the paragraph and update the manuscript (page 20-21, line 391-395):

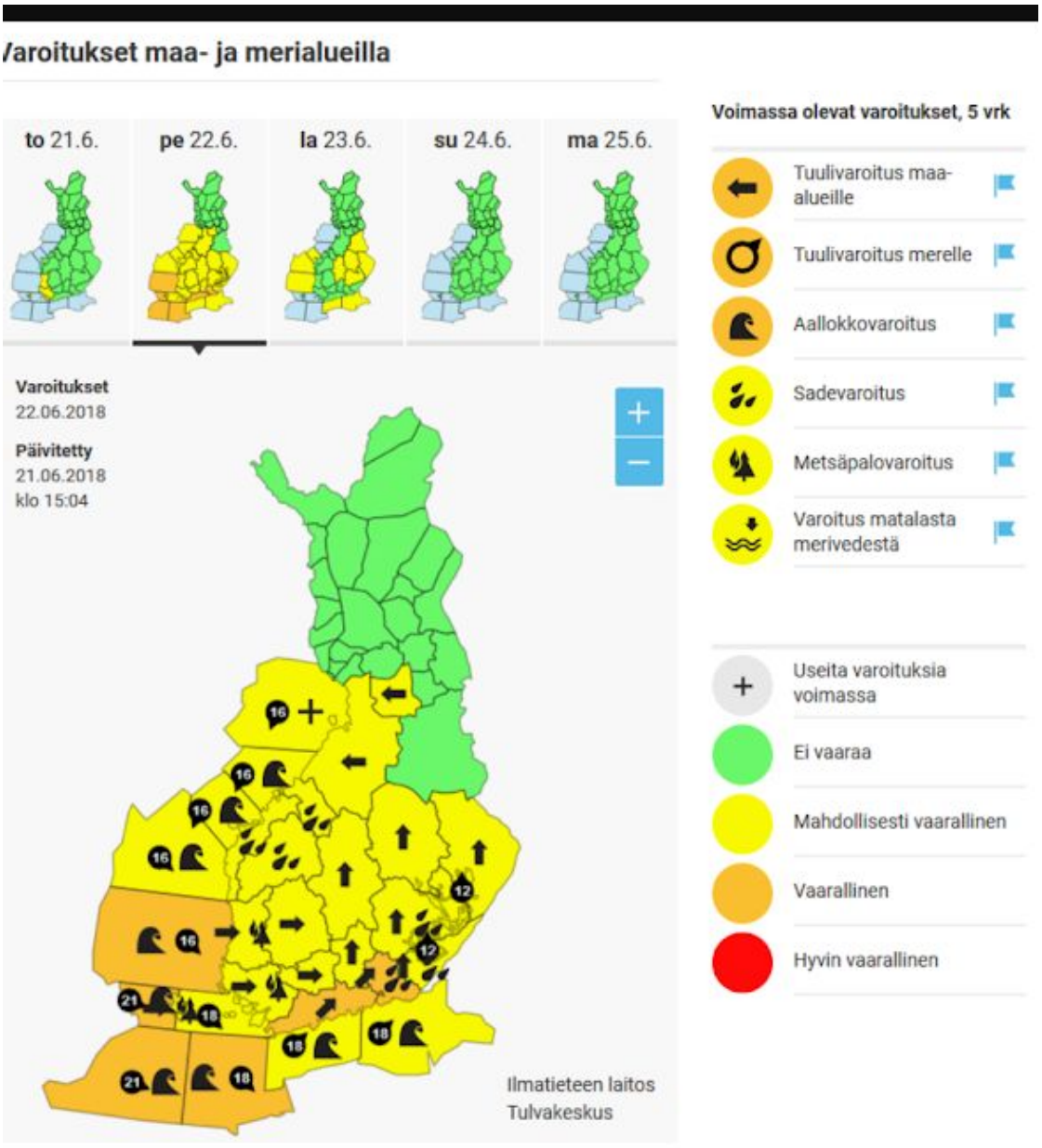
"Figure 8c presents the predicted and true damage classes at 01:00, UTC, 22 June 2018. We chose extratropical storm Pauliina as an example storm for two reasons: 1) Pauliina represents a low damage class 2) Pauliina represents a rare, summer-season extratropical storm. Figure 8c shows the predicted and true classes correlating. While weather warnings were issued to large areas in southern and middle parts of Finland, (myrskyvaroitus.com, 2018) predicted and true damage to the power grid occurred in a relatively small geographical area."

We also added the following clarification to the introduction (page 3, lines 70-74):

"[...] The ERA5 atmospheric reanalysis (European Centre for Medium-Range Weather Forecasts, 2017) provides the primary meteorological input data for this study, while the national forest inventory provided by The Natural Resources Institute Finland (Luke) is used to represent the forest conditions in the prediction. Finally, historically occurred power outages from two sources are used to train the model. However, the operational use of the model would require the use of weather prediction data instead of reanalysis."

And following clarification to the conclusion (page 22, line 406-407):

“The evaluation was, however, based on the ERA5 reanalysis data. Using the method in operations would require the use of weather prediction data, which introduces additional uncertainty to the outage prediction.”



Figures A1 and A2: The figure labels are hardly readable and the figure caption is not self-explanatory. There are abbreviations used in the figure titles which are not defined. Please spend some more words on what is shown on the figures. Can you explain the peak at -1000 in the figure titled "speed_self" and "angle_self"? It appears to be completely detached from the rest of the distribution. Why is there no blue line in the figures titled "AVG Wind gust"?

We reduced the number of shown parameters to enlarge label size. We also replaced “speed_self”, “angle_self”, “area_m2”, and “area diff” with corresponding feature names listed in Table 1. We added the following caption to the figures so that the figures should be self-explanatory:

“Histogram of and fitted Gaussian distribution of selected predictive parameters in the local dataset. The Gaussian distribution is fitted separately to all samples and samples with little outages and many outages (classes 1 and 2 specified in Section 3.3).”

Peaks at -1000 represent missing values. We dropped samples with missing values, which changed the fitted distributions a little. In particular, the differences between the mean values of the distributions reduce, which makes the deduction a little more challenging. Nevertheless, the same parameters still stand out in the analysis.

Fitted Gaussian distributions marked with the blue line have been missing in the original Figures A1 and A2 because of missing values. After dropping all samples with missing values (technically all rows having values -1000 and np.nan), the fit is successful also to AVG Wind gust, MAX Wind gust, and STD Wind gust, and mean Forest stand mean height.

Figures are updated in the manuscript and shown below.

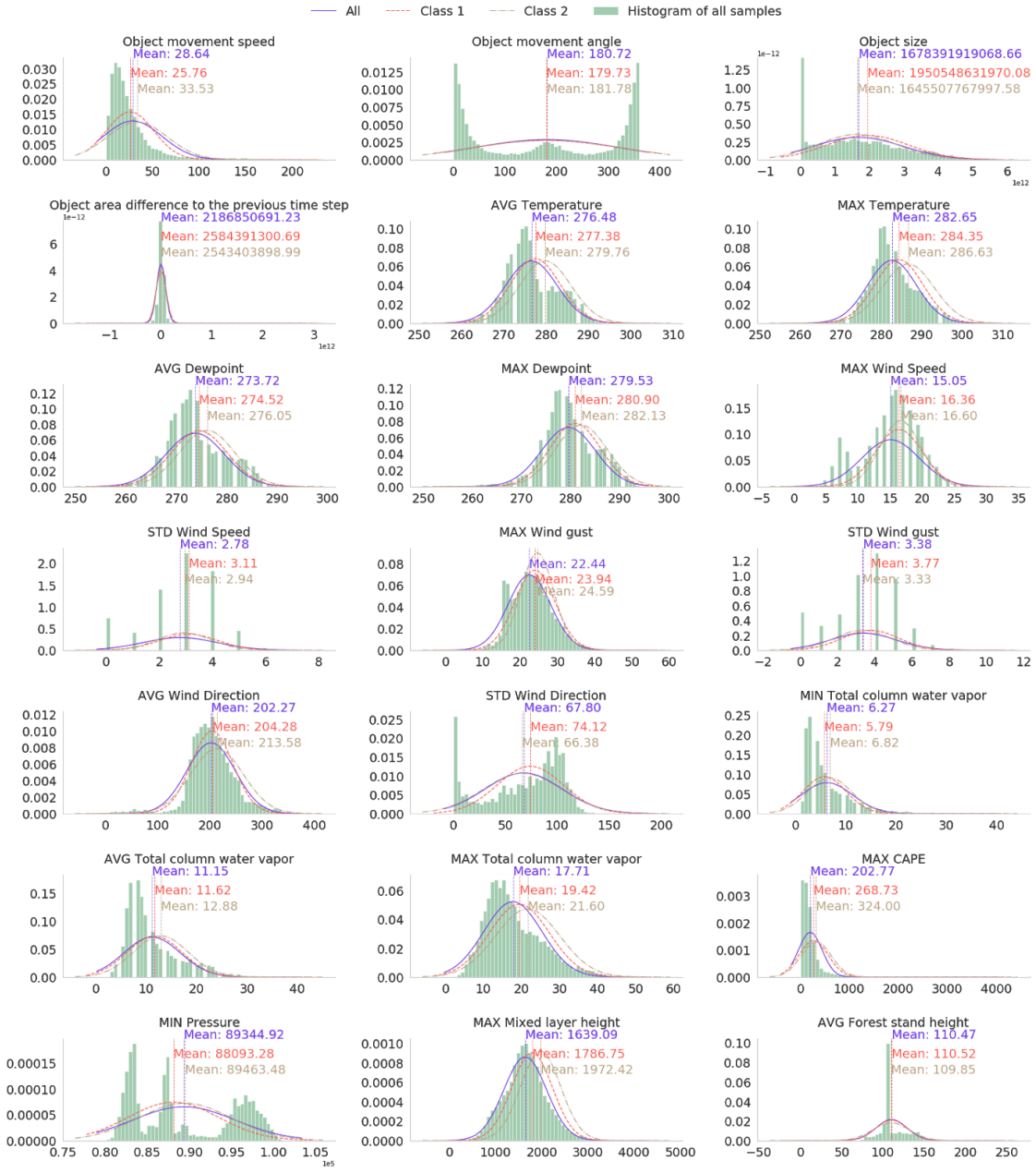


Figure A1. Histogram of and fitted Gaussian distribution of selected predictive parameters in the local dataset. The Gaussian distribution is fitted separately to all samples and samples with little outages and many outages (classes 1 and 2 specified in Section 3.3).

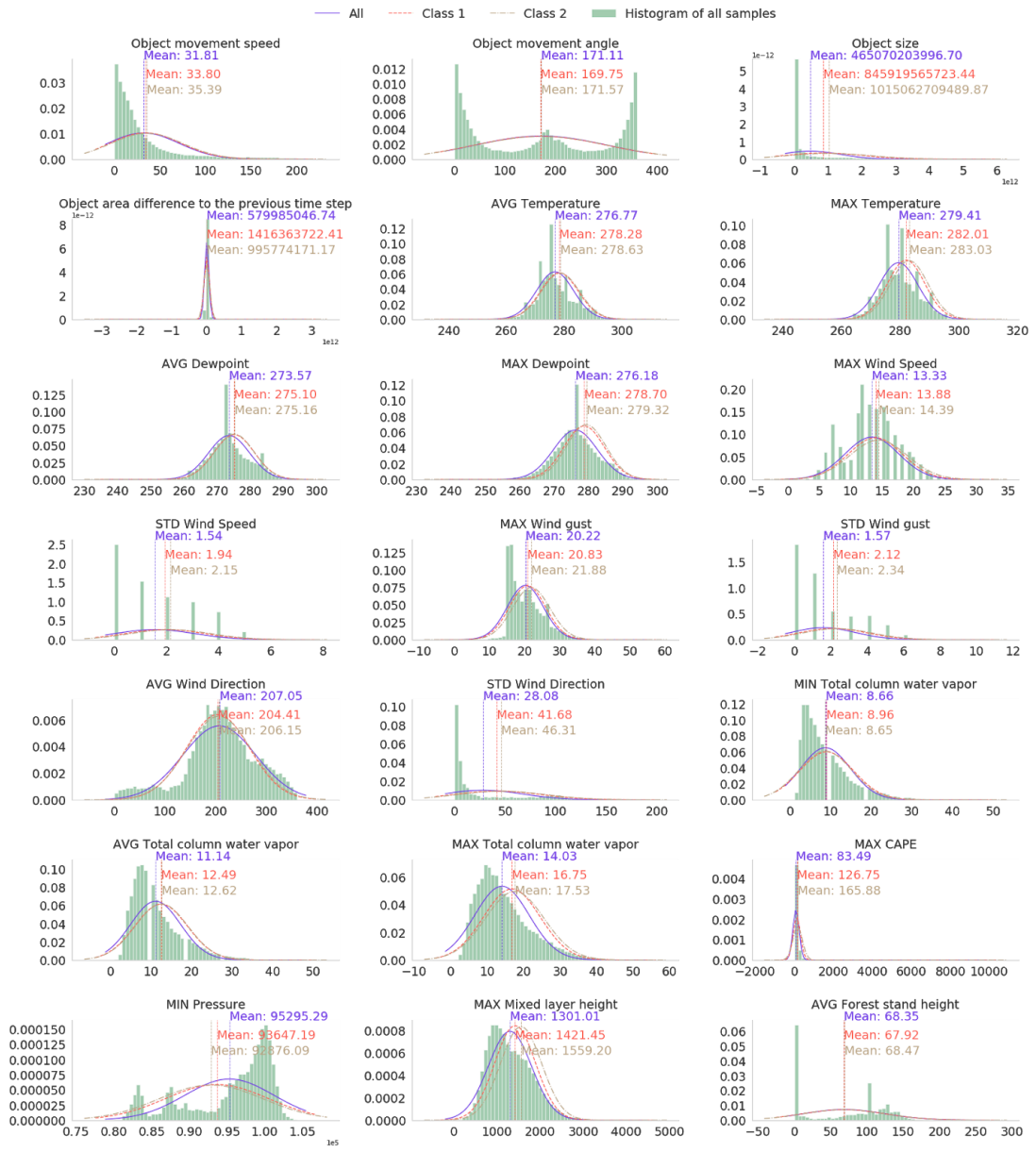


Figure A2. Histogram of and fitted Gaussian distribution of selected predictive parameters in the national dataset. The Gaussian distribution is fitted separately to all samples and samples with little outages and many outages (classes 1 and 2 specified in Section 3.3).

Technical comments:

page 2, line 50 "showed that" instead of "showed at"

We did this correction with compliments.

page 3, lines 63-66: Please check the description of the paper organization. There are missing words and incomplete sentences.

We changed and modified the paragraph as follows (Page 3, lines 75-79):

> “This paper is organized as follows: Chapter 2 presents the used data, which is followed by a step-by-step method description in Chapter 3. Chapter 3.1 discusses identifying storm objects and explains the storm tracking algorithm. Chapter 3.2 considers storm and forest characteristics, hereafter called features. Chapter 3.3 discusses how to define labels of storm objects based on the outage data. Chapter 3.4 describes the used machine learning methods. In Chapter 4, we discuss the performance of the method. Finally, Chapter 5 includes a discussion and conclusions."

page 7, line 136: Do not use blank spaces to separate numbers in order to prevent line breaks.

> We prevented line breaks in the middle of numbers using the latex \mbox command but preferred to keep spaces for clarity.

Referee 2

Tim Kruschke

Responds to the general remarks

The manuscript "Predicting power outages caused by extratropical storms" by Tervo et al. presents a novel method to predict the danger of extratropical storms to cause power outages over Finland, which is mainly due to windthrow in forest landscapes.Based on meteorological data taken from the ERA5-reanalysis as well as forest inventory data and power outage information from two local power network companies and the national responsible authority, they developed and tested classification schemes potentially suitable for warning purposes by distinguishing between severe damage events, small damage events, and no damage events. This is certainly a very interesting and relevant topic and deserves publication in NHESS. However, I consider a number of modifications necessary before publishing.

General comments:

a) A general shortcoming I notice in the prediction and its evaluation is the lack of any geographical assignment. In principle the predicted event is just "severe damage", "small damage", or "no damage" for Finland as a whole, just complemented by the polygon(s) of the storm objects. From a user-perspective (electric power network providers etc.) the question is if such a prediction is really useful facing the potential consequences, that is the alert of manpower to fix potential damages to power lines which will be rather concentrated in specific regions for most events. Of course it is better than nothing but I am sure that the method could be easily advanced to provide more regionalized information. The least thing that could have been done is to provide information on the detail level of the (power network) input data. This would mean something like "severe damage in local network 1, small damage in local network 2 and region 3 of the national network".

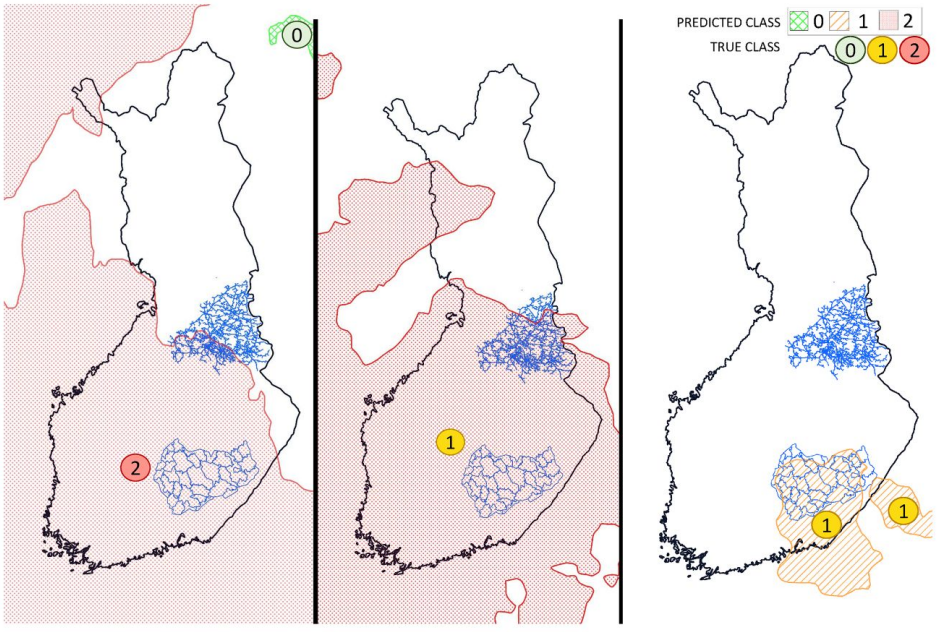
> The prediction is done for each polygon separately. Typically, such polygons cover only parts of Finland at the time. Thus, we do not predict the amount of damage to the whole of Finland, but only to the areas affected by an extratropical storm. Therefore, power grid operators could receive information about whether the storm hits the eastern or western part of the country and whether the damage in this region is expected to be light or severe. Moreover, in cases where a storm consists of several separate polygons, we are able to distinguish the damage potential of each polygon. Some examples of the coverages are illustrated in the manuscript case examples, Figure 8 (also attached below). The two first examples are extreme cases where coverage is exceptionally broad, while the third example represents a more typical geographical scale. We also attached two other examples to this response to illustrate a typical geographical scale of the prediction.

We clarified the geographical area in the introduction of the updated manuscript on page 3, lines 62-66:

“We present a novel method to identify, track, and classify extratropical storm objects based on how much power outages they are expected to induce. We adapt convective storm object detection (Rossi (2015), Tervo et al. (2019), Cintineo et al. (2014)) to find potentially harmful areas from extratropical storms by contouring objects from pressure and wind gust fields. Instead of highly-localized convective storms, we aim at larger but still regional geospatial accuracy so that, for example, damages in western and eastern Finland can be distinguished. [...]”

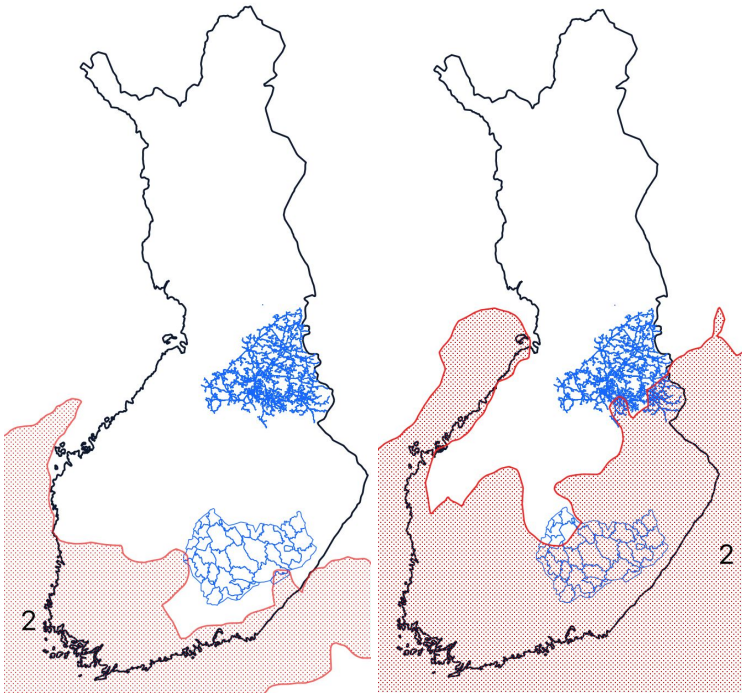
and in the chapter 3 Method on page 4, lines 116-121:

“We predict power outages by classifying storm objects identified from gridded weather data into three classes based on the number of power outages the storm typically causes. The overall process consists of the following steps: (1) identifying storm objects from weather fields by finding contour lines of particular thresholds, (2) tracking the storm object movement, (3) gathering features of the storm objects, and (4) classifying each storm object individually. The classification is conducted for each storm object separately to distinguish the different damage potential. Tracking is, however, necessary to gather necessary features such as object movement speed and direction. In the following, we discuss these phases in more detail.”



(a) Tapani, 26 December 2011 11:00 (b) Rauli, 27 August 2016 10:00 (c) Pauliina, 22 June 2018 01:00

Figure 8. Selected examples (a) Extratropical storm Tapani (b) Extratropical storm Rauli (c) Extratropical storm Pauliina produced employing SVC model trained with national dataset. The storm objects are colored based on the predicted class while the true class is stated as a colored number over the object. The time is represented as UTC time.



Left: unnamed storm, 11th September 2010 16:00 UTC,
Right: Eino, 17th November 2013 14:00 UTC

Processing polygons instead of grid data simplifies and creates a clear presentation for the end-users. This manuscript presents the potential damage areas (storm objects) on the map, where the end-user can visually inspect whether the object intersects with the power grid. It is easy to calculate in the operative user interface, for example, how many transformers are affected or even anticipated monetary losses.

A geographical aspect has indeed been omitted from the evaluation of the method. The method may work better in one region than another. However, performing a reasonable and descriptive regional evaluation is a complicated task, and we argue that it would cause more confusion than bring value. Consider, for example, an unnamed storm example on 11th September 2010, in the figure above. The polygon is correctly classified into class 2 as it caused many outages in south-western Finland. The polygon also slightly intersects south-eastern Finland. Should it be included in the eastern Finland metrics? If included, it would cause poor performance in that region since it is a class 2 polygon but still did not cause many outages in eastern Finland. If excluded, the proper ground for excluding should be selected, and the reader should be strictly aware of the ground and its consequences.

Thus, we argue that to be concise and clear, showing aggregated metrics describes the performance better than regional ones.

b) I consider the explanations of the tested classification algorithms as too short. Maybe these different methods are self-explaining for members familiar with a variety of sophisticated classification schemes and machine-learning but I think for the majority of the NHESS-readership which I assume to be with geoscientific backgrounds these methods are

hard to assess. I would like the authors to provide a little more information about the general functionality, pros & cons, and existing studies in the context of weather and climate having made use of these approaches. For some approaches like the SVC or the GP, some of this information are already given, for others this is hardly the case.

The methods used in this work are indeed standard methods. In the initially submitted manuscript, we omitted more verbose explanations to keep the text concise. The reviewer noted an excellent point about the audience. We thus extended the explanation with advantages and disadvantages along with some references to the previous studies. Nevertheless, we tried to be as brief as possible. The updated manuscript is attached below (with equations omitted).

"**Random forest classification (RFC)** is based on a random ensemble of decision trees and aggregating results from individual trees to final estimation. Trees in the ensemble are constructed with four steps: 1) use bootstrapping to generate a random sample of the data, 2) randomly select subset of features at each node, 3) determine the best split at the node using loss function, 4) grow the full tree (Breiman, 2001). RFC is good to cope with high-dimensional data. It has also been found to provide adequate performance with imbalanced data (Tervo et al., 2019; Brown and Mues, 2012) and is widely used with weather data (for example, Karthick et al. (2020); Cerrai et al. (2019); Lagerquist et al. (2017)). The method is prone to overfit, why hyperparameter-tuning is very important. Hyperparameters used in this work are listed in Table 3. We use RFC with the Gini impurity loss function.

Support Vector Classifiers (SVC) construct a hyper-plane or classification function in a high-dimensional feature space and maximize a distance between training samples and the hyperplane. The hyper-planes may be constructed with non-linear kernels such as gaussian radial basis function (RBF) (Shawe-Taylor et al., 2004) that often reform a non-linear classification problem to linear. Operating in the high-dimensional feature space without additional computational complexity makes SVC an attractive choice to extract meaningful features from a high-dimensional data set. A domain-specific expert knowledge can also be capitalized on the kernel design. On the other hand, finding the correct kernel is often a difficult task. Training SVC is a convex optimization problem meaning that it has no local minima. Depending on the kernel, a training process may, however, be a very memory-intensive process.

Suppose SVM output is assumed to be the log odds of a positive sample. In that case, one can fit a parametric model to obtain the posterior probability function and thus get probabilities for samples to belong to the particular class (Platt et al., 1999). For more details, we request the reader to consult for example Chang and Lin (2011) and Platt et al. (1999).

We implement the SVC in two phases. First, we separate class 0 (no outages) and other samples employing SVC with radial basis function (RBF), defined in Equation 1. Second, we distinguish classes 1 and 2 using SVC with a dot-product kernel defined in Equation 2 (Williams and Rasmussen, 2006). The second phase is performed only

for the samples predicted to cause outages in the first phase. The approach is similar to the often-used one-vs-one classification, where a binary classifier is fitted for each pair of classes. In our case different kernels were used for different pairs.

Gaussian Naive Bayes (GNB) (Chan et al., 1979) is a well-known and widely used method based on the Bayesian probability theory. The method assumes that all samples are independent and identically distributed (i.i.d), which does not naturally hold for the weather data. Despite the internal structure of the data, GNB is still used for weather data (for example, Kossin and Sitkowski (2009); Cintineo et al. (2014); Karthick et al. (2020)) and worth investigating in this context. The classification rule in GNB is $\hat{y} = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$, where P(y) is a frequency of class y and $P(x_i|y)$ is a likelihood of the *i*th feature assumed to be gaussian. Because of the naive i.i.d assumption, each likelihood can be estimated separately, which helps to cope with a curse of dimensionality and enable GNB to work relatively well with small datasets. On the other hand, estimating likelihoods can be done effectively and iteratively, which enables the GNB to scale to large datasets as well. As a downside, the simple method may lack expression power to perform well in a complex context.

Gaussian Processes (GP) (Rasmussen, 2003) is a non-parametric probabilistic method that interprets the observed data points as realizations of a Gaussian random process. GP is widely used for example in weather observation interpolation kriging (Holdaway, 1996). GP is a very flexible and powerful but computationally expensive method, which tends to lose its power with high-dimensional data. GP hinges on a kernel function that encodes the covariance between different data points. As a kernel, we use a product of a dot-product kernel (Equation 2) and pairwise kernel with laplacian distance (Rupp, 2015), defined in Equation 3. The kernel parameters were optimized on the training data by maximizing the log-marginal-likelihood.

Multilayer perceptrons (MLP) (Goodfellow et al., 2016) are the most basic form of artificial neural networks. Good results achieved by MLP in predicting storms (Ukkonen and Mäkelä, 2019), they are a natural choice to experiment in this work. Neural networks are very adaptive methods as they can learn a representation of the input at their hidden layers. Unlike GNB, they do not make any assumptions about the distribution of the data. As a downside, MLP requires large amounts of data, and the training process is computing-intensive. They also have a large number of hyperparameters to be optimized, including the correct network topology.

We searched the correct model parameters and network topology for local and national datasets by running multiple iterations of random search 5-fold cross-validation to obtain the best possible micro average of F1-score (defined in Chapter 4) employing Talos library (Autonomio, 2020). The final setup composes of Nadam optimizer (Dozat, 2016), random normal initializer, and relu activation function for hidden layers. Binary cross-entropy was used as a loss function. Optimal network topology varied in different datasets: For the local dataset, the best results were obtained with a network containing three hidden layers with 75, 145, and 35 neurons.

For the national dataset, the best results were obtained with a network containing three hidden layers with 75, 195, and 300 neurons. During the optimization process, the results varied between different setups from 0.6 to 0.95 in terms of F1-score.“

c) Especially for readers with a geoscientific background (as said, probably the majority of NHESS-readership) it would be interesting to read something about the relative importance of the various factors listed in Tab. 1. I understand that this may be quite different for the different classification schemes. But at least for those schemes eventually assessed to yield the best performance a qualitative summary could be listed,may mentioning the five most important factors in order of relevance.

Based on this comment and the comment given by another Referee, we conducted a permutation feature importance analysis using the Gaussian processes (GP) model and the randomly selected test set of the national dataset. The same model and data are used to produce the case examples.

The manuscript is appended with the following chapters (page 17):

“The relevance of the individual predictive features can be explored by using the permutation test, as done by Breiman (2001). First, the baseline score of the fitted model is calculated using the test set. Then each feature is randomly permuted, and the difference in the scoring function is calculated. The random permutation is repeated 30 times for each parameter, and the average of the results is used. The procedure offers information on how important the feature is to obtain good results. It should be mentioned that highly correlated features may get low importance as other features work as a proxy to the permuted feature. However, using completely independent features is not possible in weather data since weather parameters are often dependent on each other, and eliminating even the most apparent pairs from the used features impaired the results in our experiments.

We used the macro average of F1 defined in Equation 8 as a scoring function and the randomly selected test set from the national data. The relevance is shown in Figure 7. Most features show at least little relevance for the results. The first twelve features are significantly more relevant than the rest. The most important features contain at least one representative of all meteorological parameters used in training. In other words, all employed meteorological parameters are important for the prediction, while different aggregations are contributing to the "fine-tuning" of the model.

As Figure 7 shows, the most significant parameter regarding our model performance is the average wind speed. Numerous studies support our result of wind being the most important damaging factor (Virot et al., 2016; Valta et al., 2019; Jokinen et al., 2015). They are, however, highlighting the importance of maximum wind gusts instead of the average wind. Surprisingly, in our analysis, the wind gust speed does not belong to the most critical parameters. Instead, maximum mixed layer height, related to the wind gustiness, contributes crucially to the model performance. The

dependencies between predictive features might be one reason for some parameters to have a lower rank in the results.

The stand mean diameter and height are the most important features regarding the forest parameters, which corresponds to our expectations. Previous studies also state these features to influence the wind damage in forests (Pellikka and Järvenpää, 2003) and hence indirectly electricity grids. As Pellikka and Järvenpää (2003) and Suvanto et al. (2016) discuss, also the age of the forest has an impact on storm damages. However, in the feature importance test, forest age does not seem to contribute significantly to the prediction outcome.

The most important object feature is the size of the object. Object movement speed and direction did not contribute strongly to the results. However, previous studies indicate that besides the size of the impacted area, the duration of strong winds – i.e., the movement speed of the system – influences also the amount of damage (Lamb and Knud, 1991).”

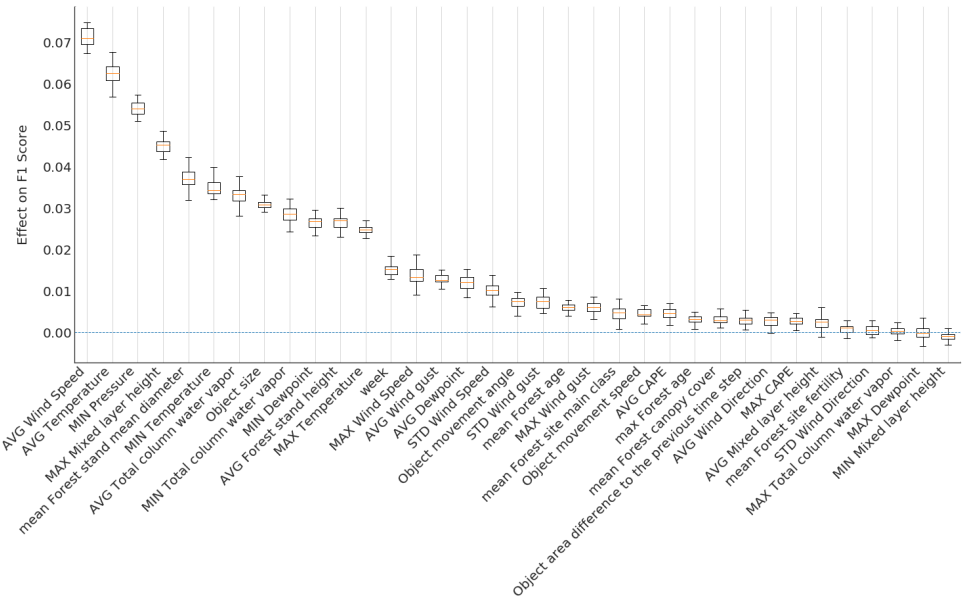


Figure7. Permutation feature importances using GP classification method trained with the randomly selected national dataset. The higher effect on the F1 score is (y-axis), the bigger is the significance.

d) As far as I understand, the evaluation metrics in equations (4)-(7) are standard metrics used in the field of machine-learning based classification. However, what I am missing in these scores is any consideration of the distance between predicted and observed class. Clearly a prediction of "severe damage" in cases of no observed damage and vice versa is worse than predicting "small damage" in these cases. But this is not reflected in any penalty for the given scores. Maybe this is a wise solution given that the classes are very different in population. Otherwise an "algorithm" always predicting no damage might be superior with respect to a score taking this distance into account. I would ask the authors at least to

comment on this matter and explain why they do not penalize larger distance between prediction and observation.

This is an important notation. The used metrics are indeed selected to take the imbalance between classes into account. We also want to provide the results in well-known standard metrics to give the reader an intuitive image of the performance. The only metrics, which take the class distance into account, we are aware of, are Gandin and Murphy Equitable Score (GMSS) and its derivatives. However, these are relatively complicated metrics and not generally known. Thus, they would hence provide only a little value to the readers.

We commented on this matter in the updated manuscript following (on page 15, line 286-289):

"[...] The selected metrics do not take a distance between predicted and true class into account. It is naturally worse to predict, for example, class 0 (no damage) in the case of true class 2 (high damage) than in the case of true class 1 (low damage). We decided, however, to use metrics that measure the method performance properly with imbalanced classes."

e) I wonder why the authors decided to provide deterministic category predictions. This Is to some degree a philosophical discussion but given the nature of the task to make a prediction and further supported by i) the rather arbitrary distinction between event classes and ii) the large number of influential factors (some of them considered in the categorization schemes but many more existing in the real world), I wonder why the authors didn't design a scheme that provides probabilities for the distinct event categories. It is often argued that end-users prefer deterministic predictions but it is clearly a fact that predictions such as produced in this study are subject to significant uncertainty. So, why not making this uncertainty transparent by providing related estimates in the form of probabilities? I do not ask the authors to redesign their whole approach but please comment on this issue. Maybe it is worth considering this as a future extension for advancement of the presented approach.

As well-argued, supplying the prediction with uncertainty information might be beneficial to some end-users if appropriately presented. Presenting the uncertainty in the correct way is also under broader discussion in meteorology due to wider use of the ensemble predictions. Providing uncertainty has, however, several challenges in this context:

1. As the referee already noted, at least the power grid operators prefer simple deterministic prediction. The simple view for the prediction is especially important in daily use where operators have only a little time to investigate the predictions.
2. The uncertainty would originate as a probabilistic prediction of the classification model, which describes the confidence of the model prediction instead of the reliability of the actual predictions. In other words, the uncertainty would not consider any sources of errors not introduced to the

model. For example, the amount of leaves in the trees significantly affects the number of caused outages, but are not considered in the prediction due to shortcomings in available data. The model could predict an incorrect class with a very high confidence as it is not aware of tree leaves at all. Providing this kind of uncertainty would be easily misinterpreted by non-expert users. Similar effects can be seen in current ensemble prediction systems when the whole ensembles cluster is biased and true values are outside the confidence interval. Therefore, we argue that the performance metrics described in this work are better guidance for the prediction uncertainty.

Having said that, especially expert users like duty forecasters would benefit from the uncertainty information. We added the following future possibility to the updated manuscript (page 22, lines 434-436):

"End users, especially expert users like duty forecasters, would benefit from the uncertainty information originating as the probabilistic prediction of the classification model. However, the presentation of such information should be very carefully chosen not to mislead non-expert users for overconfidence."

f) A very general issue is that the authors use the term prediction (and so do I in this review) but in fact the presented approach is based on atmospheric REanalysisdata, i.e. it relies on data retroactively produced from observations. I would ask the authors to rephrase respective introductory and conclusive remarks in a way that it becomes clear that this study serves as a general introduction of this approach and a proof of concept while a quasi-operational implementation at weather services or power network providers would have to be based on actual weather predictions which will introduce additional uncertainty to the final product.

The term *prediction* is widely used in the field of machine learning in the meaning of model output. In this context, it may be confused with actual weather or outage prediction, which is not our meaning. We clarified this issue in the updated manuscript introduction following (page 3, lines 70-74):

"[...] The ERA5 atmospheric reanalysis (European Centre for Medium-Range Weather Forecasts, 2017) provides the primary meteorological input data for this study, while the national forest inventory provided by The Natural Resources Institute Finland (Luke) is used to represent the forest conditions in the prediction. Finally, historically occurred power outages from two sources are used to train the model. However, the operational use of the model would require the use of weather prediction data instead of reanalysis."

And following clarification to the conclusion (page 22, line 406-407):

"The evaluation was, however, based on the ERA5 reanalysis data. Using the method in an operational setting would require weather prediction data, which introduces additional uncertainty to the outage prediction."

g) Some of the figures need optimization. Please see my respective specific comments.

We went carefully through all figures and enlarged the labels.

h) I am not a native English speaker myself, so I usually refrain from judging the language used in manuscripts written by others. However, in this example I have the strong impression that the language should be revised. A particular example are frequently missing definite and indefinite articles ("a" and "the"). Other examples can be found in my specific comments

We carefully checked the language and made corrections to the manuscript.

Respond to the specific comments

1) line 14: Please revise your citation. This is certainly no person with the family name "Re" who is cited here but an institutional citation referring to a publication by the Munich Re.

We appreciate this note. We changed the citation and also added the URL address. (Page 1, line 14)

2) lines 20-21: "...up to 69% compared to previous years". What is meant here? Is It an increase of 69% compared to previous years or a total of 69% of outages in 2011/2013v which are associated with windstorms. If the latter is the case, then please delete "compared to previous years".

When rereading this sentence, we acknowledge that it can be easily misunderstood. During the years of 2011 and 2013, the share of windstorm-induced outages was 69% of all outages. We deleted the last part of the sentence, as the referee suggested.

3) line 27: "Ulbrich et al. (2009)", not "Ulbrich et al. (Ulbrich et al., 2009)"

We corrected this with compliments. (Page 2, line 25).

4) lines 31-33: Please rephrase to make clear that this sentence contains references to studies contradicting the fore-mentioned studies and their results.

Pointing this out made us reread and clarify the entire paragraph. We reorganized it entirely in the following way. The update can be found in the updated manuscript on page 2, line 25-37:

“As Ulbrich et al. (2009) describe, there is no scientific consensus on how the occurrence and magnitude of extratropical storms will evolve in the future. Based on existing literature, the windstorm-related damages are increasing, while it remains

unclear whether this is due to the higher exposure of society or the number and intensity of extratropical storms. Gregow et al. (2017) discovered that windstorm damages had increased significantly during the previous three decades, especially in northern, central, and western Europe. Also, several other studies suggest an increase in wind-related damages in Europe (Csilléry et al. (2017); Haarsma et al. (2013); Gardiner et al. (2010)). Interestingly, some studies detected a decrease in the total number of extratropical storms (i.e. Donat et al. (2011)), while others found an increase in the number of extreme storms in specific regions, like western Europe and northeast Atlantic (Pinto et al., 2013). Another supporting view of a potential increase in extratropical storms in northern Europe can be found in the IPCC (2018) report. The report states that extratropical storm tracks are shifting towards the poles, which might affect the storminess in northern Europe. Thus, it may be concluded that also the losses related to extratropical storms are likely to increase especially in northern Europe. However, as Barredo (2010) emphasizes, the cause for increased losses can at least partly be explained by the increasing exposure of society rather than the increased number of windstorms."

5) line 46: Delete "large-scale storms" and "small-scale storms" and just name the meteorological phenomena themselves as they are now listed in brackets. It is misleading to call hurricanes large-scale if then coming to the extra-tropical storms which are even larger in spatial scale.

We changed this in the manuscript, as suggested.

6) lines 52-55: The purpose of this sentence is a little unclear to me, especially the reference to the IPCC-SREX-report. It's fine citing this report but not as one of several/many examples supporting this statement. It is basically the probably most comprehensive summary/review of studies indicating this.

This sentence and the reference is indeed detached from the previous sentences. We rephrased the end of the paragraph as follows (updated manuscript page 3, lines 58-61):

"The framework of IPCC (2018) emphasizes that the impacts of extreme weather risks can be analyzed by estimating the hazard, vulnerability, and exposure. In an increasing manner, connecting these fields (i.e. the natural hazard with the societal factors) is done with machine learning (Chen et al. (2008))."

7) line 64: Maybe replace "features" by "storm object features" or "storm object characteristics"

The features contain both storm and forest characteristics. We changed that into form (page 3, line 77):

"[...] Chapter 3.2 considers storm and forest characteristics hereafter called features. [...]"

8) lines 68-73: Please indicate the purpose of each dataset in this study, e.g. "the ERA5 atmospheric reanalysis (Hersbach et al., 2019) provides the primary meteorological input data for this study..."

Thank you. We clarified ERA5 and added additional sentences about other datasets and their roles. (Updated manuscript page 3, lines 70-74):

"The ERA5 atmospheric reanalysis (European Centre for Medium-Range Weather Forecasts, 2017) provides the primary meteorological input data for this study, while the national forest inventory provided by The Natural Resources Institute Finland (Luke) is used to represent the forest conditions in the prediction. Finally, historically occurred power outages from two sources are used to train the model. However, the operational use of the model would require the use of weather prediction data instead of reanalysis."

9) lines 74-80: Please indicate explicitly which level you use regarding the ERA wind data. I guess it is the 10m-winds but this is not said here. Additionally, you may comment on the issue regarding ERA5 surface winds which is described at https://confluence.ecmwf.int/display/CKB/ERA5%3A+large+10m+winds . As far as Ican see this does not affect this study as all problematic occasions of unrealistic high wind speed happened at geographical locations far off the study domain. Still I consider this worth mentioning as some readers may not be aware of this issue in general (so the authors could contribute to a more widespread awareness of this problem) and others may be aware of the problem but not its location and related irrelevance for this study.

The 10-meter wind gust from the surface data were used. We added this elaboration to the manuscript.

We added the following comment about the high wind speeds to the updated manuscript on page 4, line 93-95:

ERA5 data are also known to contain unrealistically large surface wind speeds in some locations (European Centre for Medium-Range Weather Forecasts, 2019). None of these locations are, nevertheless, inside the geographical domain of this work."

10) lines 84-88: What is the specific benefit of using the two local datasets on top of the national dataset for this study? Please comment.

While containing basically the same information, they also differ significantly. The national dataset contains many more outages than the local datasets, but the outages are reported with lower geographical accuracy. We train our classification models with both datasets to evaluate their performance for different types of data.

We added this information to the updated manuscript on page 4, lines 108-114. We also improved Figure 1 and moved it in the data section based on another referee's comment.

Figure 1 illustrates the geographical coverage of the power outage data. The local dataset contains all outages from 2010 to 2018 in the northern area (Loiste) and outages related to major storms in the southern area (JSE), shown in Figure 1a. The national dataset contains all outages in Finland from 2010 to 2018 divided into five regions, shown in Figure 1b. The national dataset contains in total 6 140 434 outages with relatively low geographical accuracy. On the other hand, the local dataset represents a substantially smaller geographical area with a good geographical accuracy but contains only 22 028 outages in total. We train our classification models, described in more detail in Chapter 3.4, with both datasets to evaluate their performance for different types of data."

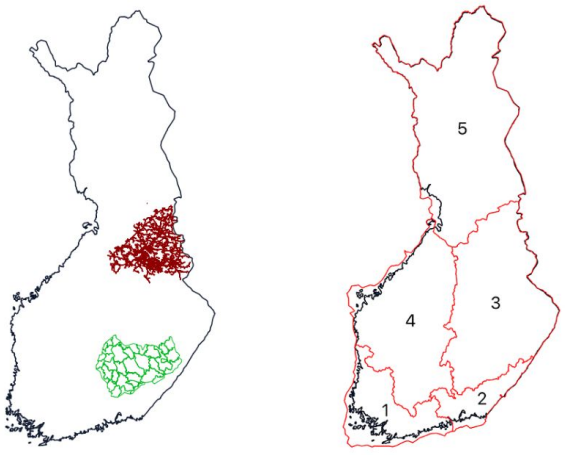


Figure 1. (a) Geographical coverage of the outage data (local dataset). The red lines represent the power grid of Loiste (northern grid company) and the green lines the operative areas of JSE (southern grid company). Outages of the local dataset are collected from both of these areas. (b) Regions in the national outage dataset. Outages are gathered from the whole Finland and aggregated to the regions shown in the figure.

11) lines 97-99: The reason given for using a threshold of 15m/s is valid as long as observed winds are considered. However, ERA5-winds are not observed winds, especially regarding gusts. It's basically model results. It should be noted here already, that at least a little bit of sensitivity tests have been performed yielding 15m/s to be the "best" choice. However, the motivation behind this study to develop a scheme which is applicable for quasi-operational forecasts would imply a transfer to a different source of meteorological data, basically weather predictions. Weather prediction models feature quite different distributions of surface wind speeds. Hence, for such an application a thorough test of the use of this threshold would be necessary. I would like to point out that there are approaches existing in the published literature on wind damage that make use of thresholds which are tied to the specific wind climatology of respective datasets, e.g. by making use of specific quantiles rather than absolute values.

As the referee noted, the chosen threshold is supported by the previous studies (Gardiner 2013, Peltola 1999, Valta 2017) and empirical knowledge of the experienced duty forecasters and power grid operators.

We are aware that the optimal threshold depends on the chosen data source, and it is also highly dependent on the time of the year and other environmental conditions. As Peltola et al. (1999) discuss, even the specific tree species are sensitive and uprooted with different wind speed thresholds. During frozen ground and leafless periods, 15 m/s wind hardly harms any trees, but during summer months, when the trees have leaves, and the soil frost does not anchor the forest to the ground, 15 m/s can be already damaging. Thus, the used threshold depends on both data source and environmental factors, and is always a compromise.

As the referee suggests, using specific quantiles would be a proficient way to determine the correct thresholds. However, with an object-based approach, the use of quantiles is not a straightforward task since the object needs to have the same absolute value inside the application domain to be a valid polygon. Therefore, the thresholds of the objects can not be always selected optimally.

We evaluated the method with a 20 m/s threshold with worse results. The evaluation is shortly mentioned in the initially submitted manuscript on page 13, line 246. However, trying out different thresholds between 15-20 m/s might yield better results. Unfortunately, this would be an intensive computing task requiring both time and budget.

We added a discussion about this matter to the updated manuscript on page 22, 408-423:

“The presented object-based approach has both advantages and disadvantages. Extracting storm objects in advance prepro- cesses the data for machine-learning techniques, such as RFC, which do not perform feature learning. It enables machine-learning methods to focus only on the relevant parts of the data. Methods not containing feature learning, such as RFC and logistic regression, have been found to outperform neural networks for forest (Hart et al., 2019) and weather data (Tervo et al., 2019). It also leads to significantly faster training times. Processing objects instead of the grid makes it also easier to track and use object attributes such as age, speed, and movement. Moreover, objects are easy to visualize, and user interfaces may be enriched with related actions such as tracking and alarms.

On the other hand, storm objects use only aggregated attributes, which may decrease the classification accuracy when predictive features vary significantly under the storm object area. Several machine-learning methods, i.e., deep neural networks, could be trained to employ those local features to gain better accuracy. Such methods could also utilize three-dimensional data. Extracting storm objects requires a fixed threshold of wind gust and pressure, which may vary depending on the characteristics of geospatial locations. Nevertheless, the previous studies

indicate the critical threshold to wind gust speed to be the same for almost entire geospatial domain of this work (Gardiner et al., 2013). Moreover, the correct threshold may vary depending on the data source. When extending the geospatial domain or changing the data source, this might become a more serious issue, and different thresholds might be needed. One possibility to determine the optimal threshold might be to use specific quantiles of the parameter values, but this would need further investigation."

12) line 103: Do you mean "...connected to objects in preceding timesteps"?

Yes, this is what we mean. We updated the manuscript on page 5, line 133-137.

13) line 103: Why do you call this "Algorithm 1" if there is no "Algorithm 2"? Why not simply calling it "Storm tracking algorithm"?

We prefer this, possibly a little clumsy, naming to be consistent with figure and table naming and to give a clear hint for the reader about the reference to the separately described algorithm (shown on another page).

14) line 103-104: Maybe I missed something but it seems to me you are not providing any information about the criterion to define/identify a "pressure object"

Please see the answer to the next comment.

15) line 103-104: You mention "the threshold" but such a threshold has not been introduced yet. This is done a few lines below. Please rephrase.

We clarified the paragraphs describing the object identification and tracking method in the updated manuscript, page 5, lines 123-137:

Storm objects are identified by finding contour lines of 10-meter wind gust fields using 15 m s−1 thresholds from the ERA5 surface level grid with a time step of 1 hour. The contouring algorithm is capable of finding interior rings of the polygons. The used wind gust fields did not, however, contain any such cases. Thus one storm object represents a solid area (polygon) where the hourly maximum wind gust exceeds 15 m s−1 during one particular hour. The threshold of 15 m s−1 is selected as different sources indicate Finland being vulnerable for windstorms and rather moderate winds (from 15 m s−1) causing damages to forests (Valta et al., 2019; Gardiner et al., 2013). Valta et al. (2019) developed a method to estimate the windstorm impacts on forests by combining the recorded forest damages from the nine most intense storms and their observed maximum inland wind gusts. According to the formula developed in the study, the inland wind gusts of 15 m s−1 alone result in forest damages of 1800 m^3. We also identify pressure objects by finding contour lines using a 1000 hPa threshold to connect potentially distant storm objects around the low-pressure center to the same storm event.

After identification, storm objects are tracked by connecting them with each other. Each storm object is first connected to nearby pressure objects from the current and preceding time steps. If pressure objects do not exist within the distance threshold, the object is connected to nearby storm objects from the current and preceding time steps. The Algorithm enables assigning each storm object to an overall event (low pressure system) and tracking the objects' movement. Algorithm 1 shows the details of the process."

16) line 111: "That means that wind objects are not assumed to move..."

Please see the answer on the next comment.

17) line 111: "45km" instead of "45km/h"; and please add "from one hourly timestep to another

We appreciate these valuable and detailed suggestions and updated the sentence to form (page 6, line 141-143):

"[...] In other words, storm objects are not assumed to move over 200 km and pressure objects over 45 km from the preceding hourly time step (Govorushko, 2011)."

The term "wind object" was also changed to "storm object" based on the comment by another Referee to be consistent.

18) line 115: "The first group is a number of object characteristics ... which are calculated ..." to the end of this sentence.

Updated with compliments on page 6, lines 147-148.

19) line 117-118: Please provide more details how you aggregate. Are the minimum/maximum/average values calculated over all grid boxes identified to belong to the storm object (i.e. exceeding 15m/s)

Yes. We clarified this to the updated manuscript on page 6, line 148-150.

"We aggregate values as a minimum, maximum, average, and standard deviation calculated over all grid cells under the object coverage to represent each parameter with one number"

20) line 118: Replace "over" by "on"

Replaced with thanks.

21) line 119: Replace "features" by "characteristics"

Replaced with thanks.

22) line 120: Replace "in the damages" by "to the damages", "support" by "complement", and "with weather parameters" by "for weather parameters"

Replaced with thanks.

23) line 121: Replace "aggregated from" by "aggregated over".

Replaced with thanks.

24) line 124: Here you mention the samples for class 1 and 2 but the class definition has not yet been introduced. This happens in the next section. Please refer to this section and include a very brief definition of the two classes in this sentence, e.g."severe damage" and "small damage"

We restructured the text to introduce classes at the end of the Chapter on page 10, line 197 (originally on page 8, line 155).

25) line 131: Now you introduce the general class definition (no damage, low damage,high damage) but again the exact definition is found at the very end of section 3.3. Additionally, the thresholds used to distinguish between the classes, especially between the two classes containing damage, seems to be completely arbitrary. AT least there is no reason given why the respective number of outages is considered to be low-damage or high damage.

We restructured the text to introduce classes on Chapter on page 10, line 197 (originally on page 8, line 155).

The thresholds used in the class definitions are discussed more in response to comment 28.

26) Fig. 1: Looking at the red lines in Fig. 1a & b I get the impression that only the lines for the northern local dataset illustrate actual power lines. The lines for the southern local dataset rather seem to be boundaries of sub-regions or so just as Fig. 1b contains region boundaries. I suggest to use different colors for different types of information. The spatial distribution of outages in Fig. 1c & d seems to having been smoothed. If so, please indicate this and the reason for doing so.

This is a valid point, and the other Referee pointed this out as well. The differences between the network topologies are simply explained by the data we have received from the two individual companies. From the northern company (Loiste), we received a shapefile of their grid. The southern company (JSE) provided their operational areas instead of the grid topology. Therefore, these two topologies look so different, even though JSE’s grid also is similar compared to Loiste.

We have now separated Figures 1a and 1b from 1c and 1d and improved the figures based on the suggestions of both referees. (Pages 5 and 9 in the updated manuscript).

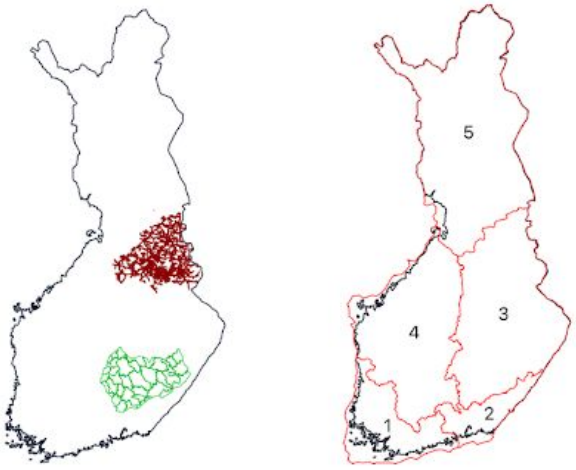


Figure 1. (a) Geographical coverage of the outage data (local dataset). The red lines represent the power grid of Loiste (northern grid company) district and the green lines the operative areas of JSE (southern grid company). Outages of the local dataset are collected from both of these areas. (b) Regions in the national outage dataset. Outages are gathered from the whole Finland but aggregated to the regions shown in the figure.

The spatial distribution of the power outages has been produced as a spatial heatmap. In other words, it is represented as a density of outages. This visualization technique is selected to illustrate the spatial distribution of a large dataset as well as possible. We updated the figure based on the other referee's comment and clarified the visualization technique in the caption.

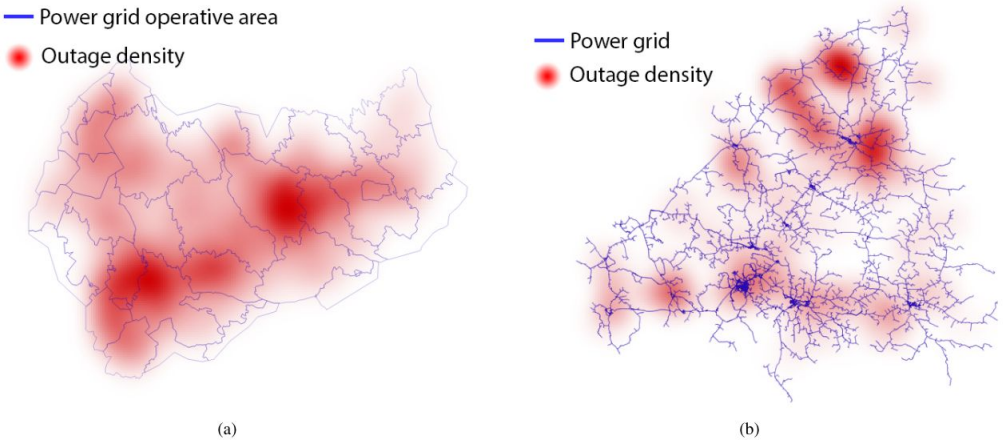


Figure 2. Spatial distribution of the outages between 2010 and 2018 visualised as a spatial heatmap. (a) JSE network (southern area) (b) Loiste network (northern area)

27) lines 143-149 (and especially when reading lines 145-146): The reader immediately wonders why the authors stay with the 15m/s-threshold and why this is not analyzed in terms

of quantitative measures. A simple example might be hit rates and false alarm rates or so. It is only in Sec. 4 (lines 248-250) that the authors write that storm identification with 15m/s yields a better basis for the following classification. Please Refer to this later explanation here.

We referred to the explanation in the updated manuscript on page 10, line 186-187.

28) lines 155-158: Eventually the class definitions seem to be set arbitrarily. If there is a reason behind the particular thresholds, please name these.

We find that when designing new tools, especially impact forecasting/estimation tools, some arbitrary "first guesses" have to be taken. As mentioned in the manuscript, the limits are designed together with the power distribution companies and duty forecasters, and they aim to be as simple and intuitive as possible. However, power grid operators do not have any specific thresholds where the actions are taken. We are also not aware of any previous studies justifying any specific thresholds, especially in Finnish conditions.

The distinction between class 1 (low damage) and class 2 (high damage) is designed so that class 2 is truly exceptional. Class 2 represents roughly 20 percent of all samples, causing at least some damage and roughly 3 percent of all samples in both datasets.

Notably, the limits can be relatively easily changed in the future based on the end-users requirements or further research.

29) lines 160-161: Why is centering and normalization necessary? Probably for some classification algorithms but not for all of them, right?

The centering and normalization are necessary for all methods except the Random Forest Classification (RFC). RFC is a decision tree method which creates the splits based on the order of the values to each feature separately. Thus, the normalization and centering do not bother RFC either.

30) lines 162-163: Please describe briefly what the application of SMOTE means and why this is beneficial/necessary.

We added the following description about the SMOTE to the manuscript on page 11-12, lines 204-211:

"[...] To cope with the imbalanced class distribution, we generate artificial training samples using the synthetic minority oversampling technique SMOTE (Chawla et al., 2002). The SMOTE creates new training samples based on their k=5 nearest neighbors following:

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i)$$

where x_i is the original sample, x_{zi} is one of x_i's k-nearest neighbour and λ is a random variable drawn uniformly from the interval [0,1]. After augmentation, all classes have an equal number of samples, which reduces classification methods' tendency to always predict the majority class."

31) lines 204-206: Why did you choose this specific topology? Did you test others? How is the sensitivity of the results to the networks topology?

The topology was searched by iterating different combinations of topologies and hyperparameters and searching for the best possible results. We clarified this into the manuscript following (page 14, lines 270-276):

"We searched the correct model parameters and network topology for local and national datasets by running multiple iterations of random search 5-fold cross-validation to obtain the best possible micro average of F1-score (defined in Chapter 4) employing Talos library (Autonomio, 2020). The final setup composes of Nadam optimizer (Dozat, 2016), random normal initializer, and relu activation function for hidden layers. Binary cross-entropy was used as a loss function. Optimal network topology varied in different datasets: For the local dataset, the best results were obtained with a network containing three hidden layers with 75, 145, and 35 neurons. For the national dataset, the best results were obtained with a network containing three hidden layers with 75, 195, and 300 neurons. During the optimization process, the results varied between different setups from 0.6 to 0.95 in terms of F1-score."

As also stated in the updated manuscript, the results varied from 0.6 to 0.95 in terms of F1-score. KDE plot of the results from the final iteration of searching the best possible network topology for the local dataset is attached below as an example.

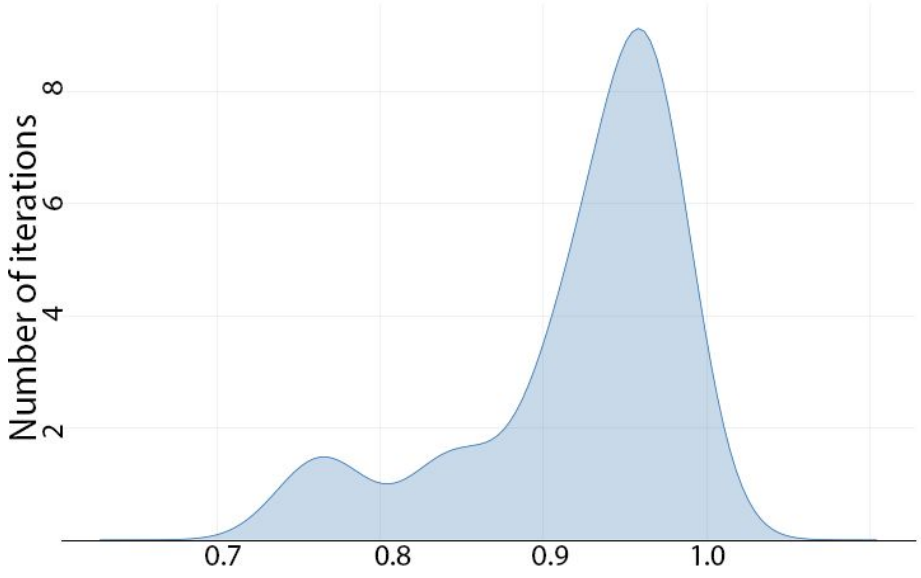


Figure: KDE plot of the results from the final iteration of hyperparameter and topology search.

32) line 236: Please explain the content of the confusion matrices briefly. Again this is probably clear to people profound in machine-learning based classification but not necessarily to the general readership in geosciences. If I understand correctly, it is simply the ratio of cases for each observed class that is show in the cells for the respective predicted classes, right?

We added the following information to the manuscript on page 18:

"Each cell of the confusion matrices represents a share of predictions having a corresponding combination of predicted and true class. For example, the middle right cell tells the share of samples belonging to class 1 but predicted to have class 2."

33) Section 4.1: This whole section is where my major comment a) becomes visible. If I understand correctly, it is just the event as a whole which is assigned with the respective category, complemented by the polygon of the storm object(s). Is it possible that different objects of one specific event are assigned with different classes? Fig.6a seems as if this is possible. On the other hand the northeastern object is outside of Finland, so it is clear that there is no damage (to Finnish power lines) observed. In this context it becomes also visible that intra-object refinement of the classification would be desirable. It makes hardly any sense for a prediction of potential damage to power lines (due to windthrow in forests) that the storm objects extend over the Baltic Sea. I understand that this is due to the primary identification being solely based on the exceedance of the wind speed threshold. However, I ask the authors to thin and comment on my general comment a). Additionally, this case study validation refers to observed wind gusts when qualitatively assessing the credibility of these specific predictions. But the authors made it very clear that the potential damage due to windstorm depends on many more factors, partly non-meteorological but related to the forests themselves. This raises again the question of relative importance of the various factors.

The classification is done to each storm object separately, and only power lines covered by the object are affected. Thus, the geographical areas can be distinguished in many cases. Furthermore, objects outside the area of coverage can be ignored.

Showing objects outside of Finland, for example, the Baltic Sea provides valuable information nevertheless to the operators in the form of preliminary information about approaching storms. The particular message in those cases is: The storm as it is now, would be (or would not be) hazardous to our power network if it was in our region. This gives the operator more tools and time to prepare.

We clarified the geographical coverage and the individual classification of storm objects in the introduction and method Chapter (please see the response to a general comment a).

We conducted a feature importance study and added it to the updated manuscript (response to general comment c).

34) lines 306-307: This sentence ignores the fact that the actual study was based on reanalysis data. Using actual weather predictions - which would be necessary for this prospect mentioned here - would introduce additional uncertainty and very likely lead to worse results than derived in the current study. This does not lower the value of the current study but it is worth mentioning when writing about such potential quasi-operational applicability.

We added the following clarification to the updated manuscript on page 22, line 406-407:

"The evaluation was, however, based on the ERA5 reanalysis data. Using the method in operations would require the use of weather prediction data, which introduces additional uncertainty to the outage prediction."

35) line 309: Start the sentence with "Including data on..."

Modified with compliments.

36) line 309: I agree that including data about forest soil and leaf index would probably be beneficial but it is questionable if such data is available in sufficient spatial and temporal resolution and coverage

The availability of such data is indeed questionable. We added a notation about this to the manuscript on page 22, lines 425-427:

"Including data on soil moisture, soil temperature, and leaf index would most probably enhance the results, if available with sufficient spatial and temporal resolution, since they would provide critical information about the environmental conditions."

37) Appendix A: All text elements and axis labels in figures A1 and A2 are hardly readable.

We reduced the number of shown parameters to enlarge label size. We also replaced "speed_self", "angle_self", "area_m2", and "area diff" with corresponding feature names listed in Table 1. The updated figures are attached below:

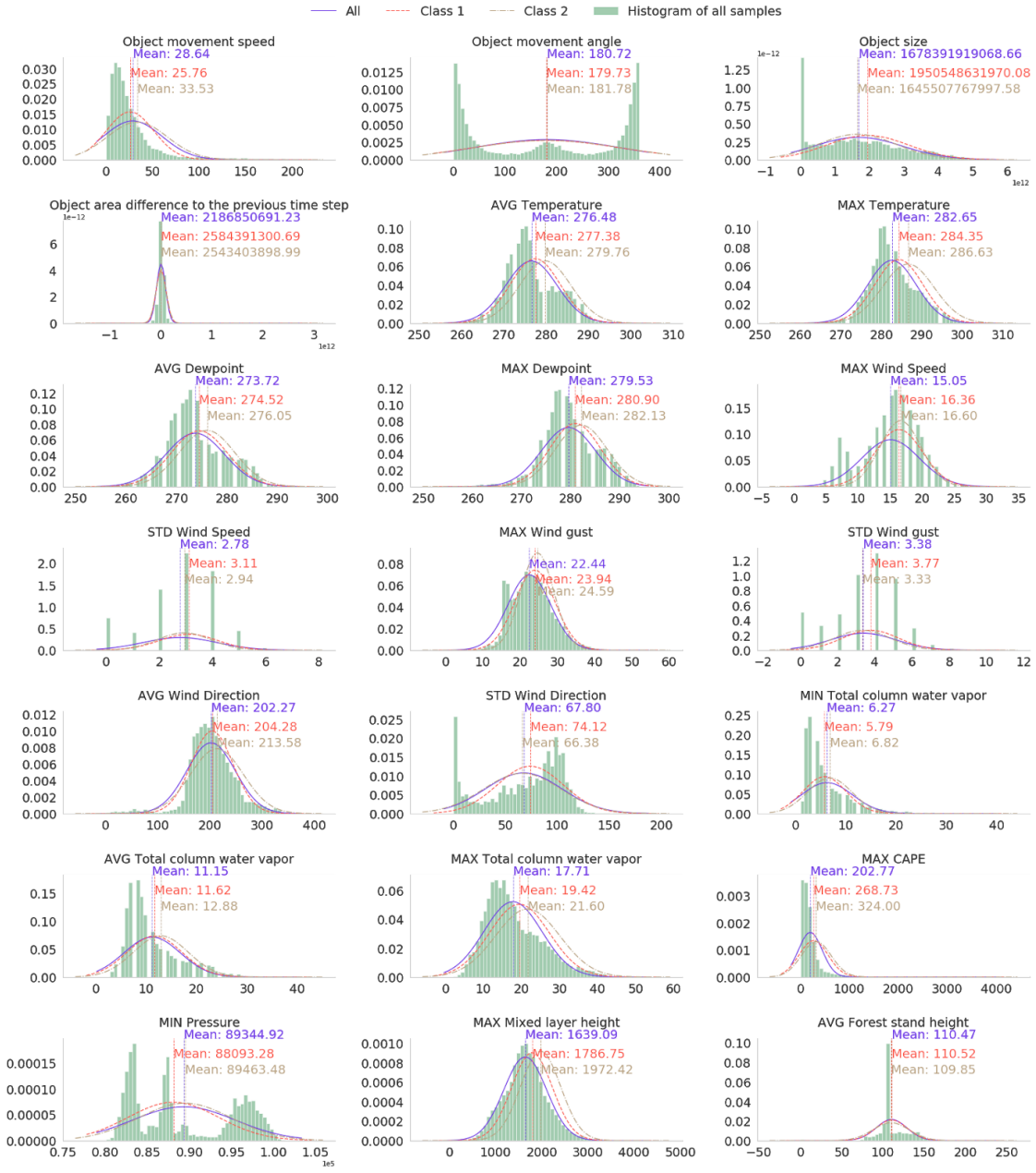


Figure A1. Histogram of and fitted Gaussian distribution of selected predictive parameters in the local dataset. The Gaussian distribution is fitted separately to all samples and samples with little outages and many outages (classes 1 and 2 specified in Section 3.3).

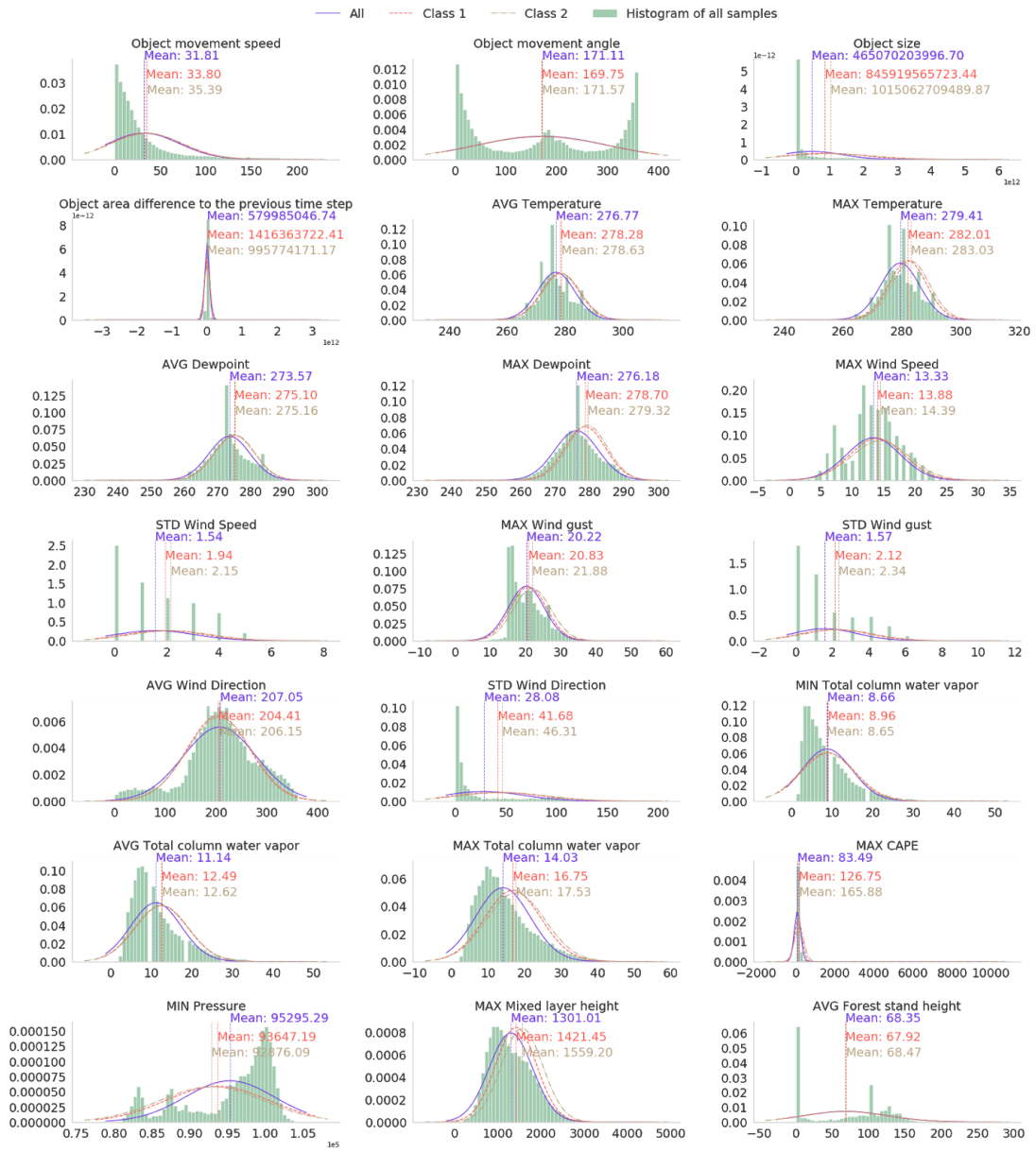


Figure A2. Histogram of and fitted Gaussian distribution of selected predictive parameters in the national dataset. The Gaussian distribution is fitted separately to all samples and samples with little outages and many outages (classes 1 and 2 specified in Section 3.3).

Short comment 1

Peter Watson

Responds to the general remarks

A thoroughly interesting paper. The methodology for identifying storms is especially interesting. However, there may be a few ways to improve the work presented. More specifically:

1) In lines 46 to 48, the authors claim that modeling power outages caused by extratropical events is an understudied problem. However there are actually several papers that describe a power outage prediction system designed specifically for modeling power outages from extratropical storms that are not cited: Yang et al, https://www.mdpi.com/2071-1050/12/4/1525; and Cerrai et al, https://ieeexplore.ieee.org/abstract/document/8656482

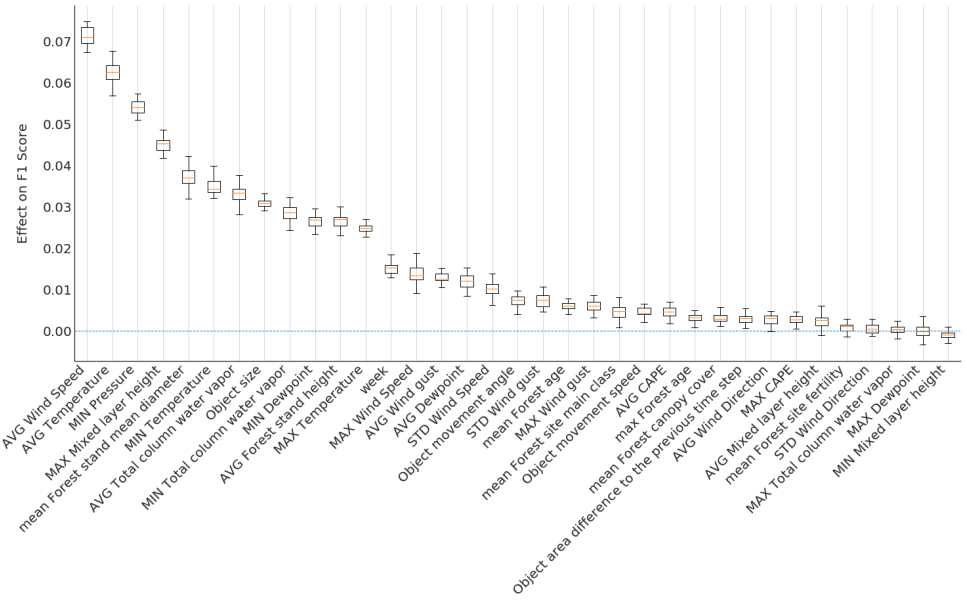
> We appreciate this advice and added mentioned papers to the previous work. (Updated manuscript page 2, lines 52-54).

2) In figure 4b, it's unclear why the data contains prominent examples where there are very few or no outages, but have a large number of customers affected. Is this trendreal, or is it an artifact of noise in the data?

> Only six customers usually lose electricity in one outage in the national datasets. In some cases, however, outages affect many more customers. We can not ensure the correctness of each data point, but we did check some extreme cases. Typically these cases occur in urban areas and are rare because the power network is mainly underground in these areas.
>
> We added a comment about this matter to the updated manuscript in page 10, lines 195-196:
>
> "Notably, in some rare cases, many more customers are affected. Based on our random inspections, these cases occur typically in urban areas and are rare because the power network is mainly underground in these areas."

3) By using week as a predictor variable the authors may be over-fitting. For example,to my knowledge, there's no specific mechanism of why a storm on the 42nd week of the year would be particularly strong. But if you had several examples of strong storms on that week, the model would learn that trend and begin to predict strong outages just because of the week, independent of the actual meteorological characteristics of the storms. There are probably other, less problematic ways to describe seasonal aspects of storms to the model.

This is truly a very valid concern. During the review process, we conducted a permutation feature importance analysis using the Gaussian processes (GP) model and randomly selected test set of the national dataset. The results of the analysis are shown below. Please consult responses to RC1 or RC2 for more information.



The results indicate that permuting week during the training process had only a little effect (0.015 +/- 0.001) on F1 score. Moreover, we conducted the feature importance study also using corresponding train data. In that case the week had almost the same effect (0.013 +/- 0.002) on the F1 score, which also indicates that using the week as a predictor does not lead to overfitting.

4) I would recommend a more rigorous and comprehensive method for validating the model. As discussed in the paper, the k-fold cross-validation approach may not sufficiently isolate temporally or spatially correlated information from the model, and thus inflate the model's performance. The 2010 to 2011 holdout approach is presented as an alternative to this approach, but the types of storm events that occur often vary widely from year to year. A leave-one-day/week/month/year-out cross validation (where foreach day, week, month, or year in the database you hold out that data, train the model on the remaining data, and predict on the withheld data. Then evaluate the model on all of those results) would provide more compelling results.

Thank you for the comment.

First, we would like to clarify that the test results does not represent k-fold cross-validation but randomly selected holdout as stated in the beginning of Chapter 4 (page 12, lines 208-209 in the originally submitted manuscript).

The validation could still be extended to more rigorous methods like the one suggested by the reviewer. We are aware of a potential autocorrelation issue when selecting the testset randomly. We selected to address this issue by solid year

holdout since based on the data analysis (for example Figure 2, attached also below) 2010 to 2011 represents the whole data relatively well in terms of number of outages and storm objects.

We would also like to note that there is no significant difference between two different testset (randomly picked and continuous holdout). Thus, we believe that our method provided sufficient validation scores.

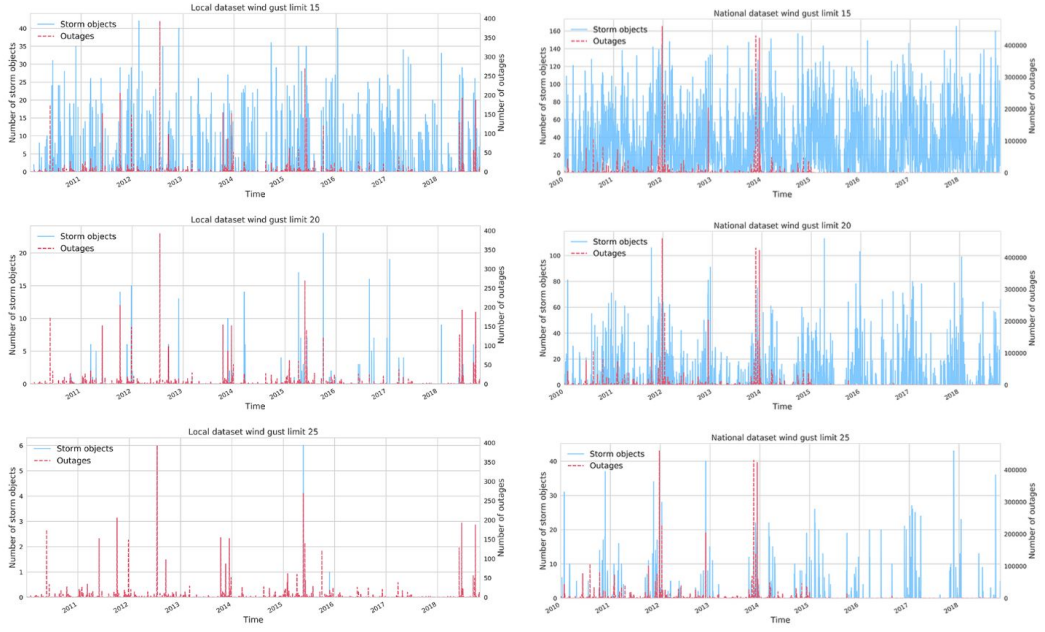


Figure 2. Storm object time series (15, 20 and 25 m s^{-1} contours) with occurred outages for local and national datasets.

Nevertheless, we commented the issue in the Discussion section (page 22, lines 430-433 in the updated manuscript) :

“Especially in the randomly selected test set, data may be autocorrelated, which may lead to unrealistically good results. We have addressed this issue by also using a continuous time series (from 2010 to 2011) for the test set. The evaluation could also be extended with a leave-one-day-out or leave-one-week-out method where for each week one day or for each month one week is hold out for validation purposes.”

Predicting power outages caused by extratropical storms

Roope Tervo[1,*], Ilona Láng[1,*], Alexander Jung[2], and Antti Mäkelä[1]

[1]Finnish Meteorological Institute, B.O. 503, 00101 Helsinki, Finland
[2]Aalto University, Dept of Computer Science, B.O. 11000, 00076 Aalto, Finland
[*]These authors contributed equally to this work.

Correspondence: Roope Tervo (roope.tervo@fmi.fi)

Abstract. Strong winds induced by extratropical storms cause a large number of power outages, especially in highly forested countries such as Finland. Thus, predicting the impact of the storms is one of the key challenges for power grid operators. This article introduces a novel method to predict the storm severity for the power grid employing ERA5 reanalysis data combined with forest inventory. We start by identifying storm objects from wind gust and pressure fields by using contour lines of 15 $\mathrm{m\,s^{-1}}$ and 1000 hPa respectively. The storm objects are then tracked and characterized with features derived from surface weather parameters and forest vegetation information. Finally, objects are classified with a supervised machine learning method based on how much damage to the power grid they are expected to cause. Random Forest Classifier, Support Vector Classifier, Naive Bayes, Gaussian Processes, and Multilayer Perceptron were evaluated for the classification task, Support Vector Classifier providing the best results.

1 Introduction

Strong winds caused by extratropical storms are among the most significant natural hazards in Europe, causing massive damage to the forests and society (e.g. Schelhaas et al. (2003); Schelhaas (2008); Ulbrich et al. (2008); Seidl et al. (2014); Valta et al. (2019)); extratropical storms are responsible for 53 percent of all losses related to natural hazards in Europe (Kron W., Schuck A., 2013). Such storms pose a huge challenge for power distribution companies in highly-forested countries such as Finland (Gardiner et al., 2010) where falling trees cause power outages for hundreds of thousands of customers every year (Niemelä, 2018). The windstorms create a significant risk for the power supply in Finland, which has over 90 000 kilometers of overhead lines (70 percent of it medium-voltage, 1-35 kV, network) passing through forest (Kufeoglu and Lehtonen, 2015). Between the years 2010 and 2018, on average 46 percent of all transmission faults in Finland were caused by extratropical storms (Finnish Energy, 2010-2018). During the years of the most damaging storms, 2011 and 2013, the share of windstorm damages of all fault causes was up to 69 percent (Finnish Energy, 2011, 2013). The need for managing power interruptions is even more urgent since the power suppliers in Finland are obliged to financially compensate customers of urban areas after 6 hours and

rural areas after 36 hours of interruption in electricity distribution (Nurmi et al., 2019). Thus they require a large amount of workforce to fix caused damages rapidly.

As Ulbrich et al. (2009) describe, there is no scientific consensus on how the occurrence and magnitude of extratropical storms will evolve in the future. Based on existing literature, the windstorm-related damages are increasing, while it remains unclear whether this is due to the higher exposure of society or the number and intensity of extratropical storms. Gregow et al. (2017) discovered that windstorm damages had increased significantly during the previous three decades, especially in northern, central, and western Europe. Also, several other studies suggest an increase in wind-related damages in Europe (Csilléry et al. (2017); Haarsma et al. (2013); Gardiner et al. (2010)). Interestingly, some studies detected a decrease in the total number of extratropical storms (i.e. Donat et al. (2011)), while others found an increase in the number of extreme storms in specific regions, like western Europe and northeast Atlantic (Pinto et al., 2013). Another supporting view of a potential increase in extratropical storms in northern Europe can be found in the IPCC (2018) report. The report states that extratropical storm tracks are shifting towards the poles, which might affect the storminess in northern Europe. Thus, it may be concluded that also the losses related to extratropical storms are likely to increase especially in northern Europe. However, as Barredo (2010) emphasizes, the cause for increased losses can at least partly be explained by the increasing exposure of society rather than the increased number of windstorms.

Several previous studies respond to the demand for storm impact estimation for power distribution, many of them focusing on the hurricane-induced power blackouts in northern America (Eskandarpour and Khodaei (2017); Guikema et al. (2014, 2010); Nateghi et al. (2014); Han et al. (2009); Wang et al. (2017); Allen et al. (2014); Chen and Kezunovic (2016); He et al. (2017); Liu et al. (2018)). Convective thunderstorms have also been investigated thoroughly. Li et al. (2015) introduced an area-based outage prediction method further developed to take power grid topology into account (Singhee and Wang, 2017). Shield et al. (2018) studied outage prediction by applying a random forest classifier to weather forecast data in a regular grid. Kankanala et al. used data from ground observation stations and experimented regression (Kankanala et al., 2011), a multilayer perceptron neural network (Kankanala et al., 2012), and ensemble learning (Kankanala et al., 2014) to predict outages caused by wind and thunder. The Bayesian outage probability (BOP) prediction model developed by Yue et al. (2018) combines weather radar data and unifies it to a regular grid. Cintineo et al. (2014) create spatial objects from satellite and weather radar data, and track and classify the objects with the Naïve Bayesian classifier. Rossi (2015) developed a method to detect and track convective storms. The method was further developed to predict power outages (Tervo et al., 2019).

While much work exists on damage caused by hurricanes and convective thunderstorms, relatively few examples exist relating to outages caused by mid-latitude extratropical storms differing from hurricanes and convective storms in available data, time-span, and applicable methods for detecting and tracking. Extratropical storms are considered, for example, in Yang et al. (2020), where different decision tree methods are applied to a regular grid in the outage prediction task. Cerrai et al. (2019) also uses decision trees and regular grid for the outage prediction taking tree-leaf conditions into account as a predictive feature. Related forest damage studies have been conducted with random forest classifiers and neural networks. Hart et al. (2019) showed that random forest regression and artificial neural networks could predict the number of falling trees in France caused by the wind. Hanewinkel (2005) conducted a similar study in Germany using artificial neural networks. Artificial

neural networks have been used to predict extreme weather in Finland (Ukkonen et al., 2017; Ukkonen and Mäkelä, 2019). The framework of IPCC (2018) emphasizes that the impacts of extreme weather risks can be analyzed by estimating the hazard, vulnerability, and exposure. In an increasing manner, connecting these fields (i.e., the natural hazard with the societal factors) is done with machine learning (Chen et al., 2008).

We present a novel method to identify, track, and classify extratropical storm objects based on how much power outages they are expected to induce. We adapt convective storm object detection (Rossi (2015), Tervo et al. (2019), Cintineo et al. (2014)) to find potentially harmful areas from extratropical storms by contouring objects from pressure and wind gust fields. Instead of highly-localized convective storms, we aim at larger but still regional geospatial accuracy so that, for example, damages in western and eastern Finland can be distinguished. We train a supervised machine learning model to classify storm objects according to their damage potential. To our knowledge, our method is the first that employs the extratropical storm objects as polygons and combines them with meteorological and non-meteorological features to predict power outages. The method can be used as a decision support tool in power distribution companies or as part of elaborating impact forecast by duty forecasters in national hydro-meteorological centers. The ERA5 atmospheric reanalysis (European Centre for Medium-Range Weather Forecasts, 2017) provides the primary meteorological input data for this study, while the national forest inventory provided by The Natural Resources Institute Finland (Luke) is used to represent the forest conditions in the prediction. Finally, historically occurred power outages from two sources are used to train the model. However, the operational use of the model would require the use of weather prediction data instead of reanalysis.

This paper is organized as follows: Chapter 2 presents the used data, which is followed by a step-by-step method description in Chapter 3. Chapter 3.1 discusses identifying storm objects and explains the storm tracking algorithm. Chapter 3.2 considers storm and forest characteristics, hereafter called features. Chapter 3.3 discusses how to define labels of storm objects based on the outage data. Chapter 3.4 describes the used machine learning methods. In Chapter 4, we discuss the performance of the method. Finally, Chapter 5 includes a discussion and conclusions.

2 Data

We base our method on three main data sources: ERA5 reanalysis data (Hersbach et al., 2019), multi-source national forest inventory (ms-nfi) provided by The Natural Resources Institute Finland (Luke), and occurred power outages obtained from two sources: First, the *local dataset* is gathered from two power distribution companies, Loiste and Järvi-Suomen Energia (JSE), located in Eastern Finland. Second, the *national dataset* is obtained from Finnish Energy (ET), a branch organization for the industrial and labor market policy of the energy sector. All data consider years from 2010 to 2018.

ERA5 is the newest generation reanalysis data provided by ECMWF. ERA5 covers the years from 1979 onward with a one-hour temporal resolution, has a horizontal resolution of 31 km, and covers the atmosphere using 137 levels up to a height of 80 km (Hersbach et al., 2019). Compared to in-situ wind observations, reanalysis data provides a spatiotemporally wider dataset. However, a question may arise about the accuracy of the reanalysis data. Ramon et al. (2019) examined the wind speed characteristics of a total of five state-of-the-art global reanalyses concerning 77 instrumented towers. In their study, ERA5

had the best agreement with in-situ observations on daily time scales; this suggests the ERA5 wind parameters to be adequate in windstorm damage examinations as well. ERA5 data are also known to contain unrealistically large surface wind speeds in some locations (European Centre for Medium-Range Weather Forecasts, 2019). None of these locations are, nevertheless, inside the geographical domain of this work.

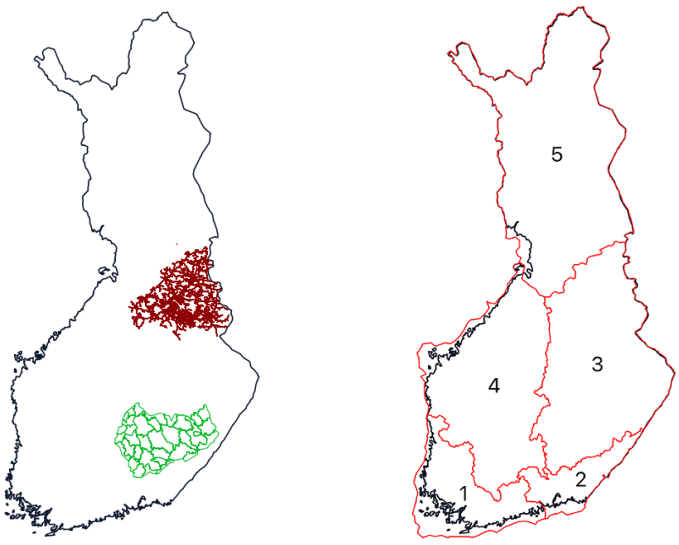
The multi-source forest inventory data is based on field measurements, satellite observations, digital maps, and other geo-referenced data sources (Mäkisara et al., 2016). The data consists of estimates for the forest age, tree species dominance, the mean and total volume, and the biomass (total and tree species-specific). The original geospatial resolution of the data is 16 meters, which has been reduced to approximately 1.6 km resolution to speed up the processing. Taking into account the size of extratropical cyclones (diameter 1000 km) and the wide areas where wind damages typically occur e.g. near to the cold front, we consider a resolution of 1.6 km being sufficiently high for modeling wind storm damages.

Power outage data are obtained from two complementary sources. *The national dataset* is acquired from the Finnish Energy (2010-2018) who aggregates the data from power distribution companies in Finland. The national data are provided only for research purposes and for areas containing a minimum of six grid companies; this is, for example, to ensure energy users' anonymity. Therefore, the national dataset does not include exact locations of the faults. We have also obtained some parts of the data with better spatial accuracy from two individual power distribution companies. In this paper, we refer to this data as *the local dataset*. In the local dataset, the fault locations are reported in relation to transformers, i.e. the spatial resolution of the outages ranges from a few meters to kilometers.

Figure 1 illustrates the geographical coverage of the power outage data. The local dataset contains all outages from 2010 to 2018 in the northern area (Loiste) and outages related to major storms in the southern area (JSE), shown in Figure 1a. The national dataset contains all outages in Finland from 2010 to 2018 divided into five regions, shown in Figure 1b. The national dataset contains in total 6 140 434 outages with relatively low geographical accuracy. On the other hand, the local dataset represents a substantially smaller geographical area with a good geographical accuracy but contains only 22 028 outages in total. We train our classification models, described in more detail in Chapter 3.4, with both datasets to evaluate their performance for different types of data.

3 Method

We predict power outages by classifying storm objects identified from gridded weather data into three classes based on the number of power outages the storm typically causes. The overall process consists of the following steps: (1) identifying storm objects from weather fields by finding contour lines of particular thresholds, (2) tracking the storm object movement, (3) gathering features of the storm objects, and (4) classifying each storm object individually. The classification is conducted for each storm object separately to distinguish the different damage potential. Tracking is, however, necessary to gather necessary features such as object movement speed and direction. In the following, we discuss these phases in more detail.



(a) (b)

Figure 1. (a) Geographical coverage of the outage data (local dataset). The red lines represent the power grid of Loiste (northern grid company) and the green lines the operative areas of JSE (southern grid company). Outages of the local dataset are collected from both areas. (b) Regions in the national outage dataset. Outages are gathered from entire Finland and aggregated to the regions shown in the figure.

3.1 Identifying and tracking storm objects

Storm objects are identified by finding contour lines of 10-meter wind gust fields using 15 $\mathrm{m\,s^{-1}}$ thresholds from the ERA5 surface level grid with a time step of 1 hour. The contouring algorithm is capable of finding interior rings of the polygons. The used wind gust fields did not, however, contain such cases. Thus one storm object represents a solid area (polygon) where the hourly maximum wind gust exceeds 15 $\mathrm{m\,s^{-1}}$ during one particular hour. The threshold of 15 $\mathrm{m\,s^{-1}}$ is selected as different sources indicate Finland being vulnerable for windstorms and rather moderate winds (from 15 $\mathrm{m\,s^{-1}}$) causing damages to forests (Valta et al., 2019; Gardiner et al., 2013). Valta et al. (2019) developed a method to estimate the windstorm impacts on forests by combining the recorded forest damages from the nine most intense storms and their observed maximum inland wind gusts. According to the formula developed in the study, the inland wind gusts of 15 $\mathrm{m\,s^{-1}}$ alone result in forest damages of 1800 $\mathrm{m^3}$. We also identify pressure objects by finding contour lines using a 1000 hPa threshold to connect potentially distant storm objects around the low-pressure center to the same storm event.

After identification, storm objects are tracked by connecting them with each other. Each storm object is first connected to nearby pressure objects from the current and preceding time steps. If pressure objects do not exist within the distance threshold, the object is connected to nearby storm objects from the current and preceding time steps. The Algorithm enables assigning each storm object to an overall event (low pressure system) and tracking the objects' movement. Algorithm 1 shows the details of the process.

We use a 500 km distance threshold for the distance between the storm and pressure objects. As the typical diameter of an extratropical storm is approximately 1000 km (Govorushko, 2011), we assume the damaging storm objects to situate a maximum 500 km from the center of the low pressure. The threshold for movement speed is 200 $\mathrm{km\ h^{-1}}$ for storm objects and 45 $\mathrm{km\ h^{-1}}$ for pressure objects. In other words, storm objects are not assumed to move more than 200 km and pressure objects more than 45 km from the preceding hourly time step (Govorushko, 2011). Convective storms may move faster but are outside the focus of this work.

3.2 Extracting storm object features

We characterize the storm objects identified by the methods discussed in Section 3.1 using the features listed in Table 1. The features are structured as four groups. The first group is a number of object characteristics such as size and movement speed and direction, which are calculated from the contoured storm objects themselves. As the second group, relevant weather conditions, such as wind speed, temperature, and others, are extracted from ERA5 data. We aggregate values as a minimum, maximum, average, and standard deviation calculated over all grid cells under the object coverage to represent each parameter with one number. Third, as most of the outages are caused by the trees falling on power grid lines (Campbell and Lowry, 2012), the characteristics of the forest contribute to the damages (Peltola et al., 1999), we complement our data with forest information. As for weather parameters, values are aggregated over the storm object coverage. The fourth group consists of the number of outages and affected customers used as labels in the model training process discussed in more detail in Chapter 3.4.

We selected the 35 parameters based on two main criteria: First, we prepared a list of potential parameters detected in related studies, e.g. Suvanto et al. (2016); Peltola et al. (1999); Valta et al. (2019), or identified through the empirical experience of duty forecasters (Weather and Safety Center of Finnish Meteorological Institute - Duty forecasters, 05/2020). Second, we selected the relevant parameters, which were available to us or accessible with a reasonable effort. However, some possibly essential parameters, like soil temperature from ERA5 reanalysis, were left out because of the slow downloading process.

After the preliminary selection of the parameters, we conducted dozens of light experiments using different combinations of parameters and models to find the best possible setup. To this end, we fitted the Gaussian distribution to each parameter using at first all samples, then samples with few outages, and finally with many outages (classes 1 and 2 specified in Section 3.3). While many other distributions are known to suit better in modeling particular parameters, such as Gamma in precipitation, Weibull in wind speed, and Lognormal in cloud properties (Wilks, 2011), the Gaussian distribution is a sufficient simplification to help in selecting relevant parameters. We visually inspected the differences between fitted Gaussian distributions to deduce the potential relevance of the parameter. Supposedly the distribution of one parameter is different for all samples and samples with many outages, and the classification method may exploit the parameter to predict the damage potential of the storm object. The distributions of some selected parameters are shown in Appendix A. In total, 35 parameters, shown as boldfaced in Table 1, were chosen for the final classification.

Algorithm 1 Storm tracking

Input

Storm and pressure objects S_o arranged by time

$pressure\ distance\ threshold$

$wind\ distance\ threshold$

$speed\ threshold$

$time\ step$

Output

Connected storm and pressure objects with storm ID

for all storm and pressure object $O_{w|p} \in S_o$ **do**

$current\ time \leftarrow$ time of the object $O_{w|p}$

$previous\ time \leftarrow current\ time - time\ step$

Current time pressure objects $S_p^c \leftarrow$ pressure objects having centroid within $pressure\ distance\ threshold$ from object $O_{w|p}$ centroid and time stamp $current\ time$

Previous time pressure objects $S_p^p \leftarrow$ pressure objects having centroid within $speed\ threshold$ from object $O_{w|p}$ centroid and time stamp $previous\ time$

Current time storm objects $S_w^c \leftarrow$ storm objects having centroid within $wind\ distance\ threshold$ from object $O_{w|p}$ centroid and time stamp $current\ time$

Previous time storm objects $S_w^p \leftarrow$ storm objects having centroid within $speed\ threshold$ from object $O_{w|p}$ centroid and time stamp $previous\ time$

if pressure object $O_p^c \in S_p^c$ exists with ID **then**

Use pressure object $O_p^c\ ID$

else if pressure object $O_p^p \in S_p^p$ exists with ID **then**

Use previous time pressure object $O_p^p\ ID$

else if storm object $O_w^c \in S_w^c$ exists with ID **then**

Use storm object $O_w^c\ ID$

else if storm object $O_w^p \in S_w^p$ exists with ID **then**

Use previous time storm object $O_w^p\ ID$

else if storm or pressure object $O_{w|p}^p \in S_w^p \cup S_p^p$ exists without ID **then**

Give new ID to the previous object $O_{w|p}^p$ and current object $O_{w|p}$

else

Leave object $O_{w|p}^p$ without ID

end if

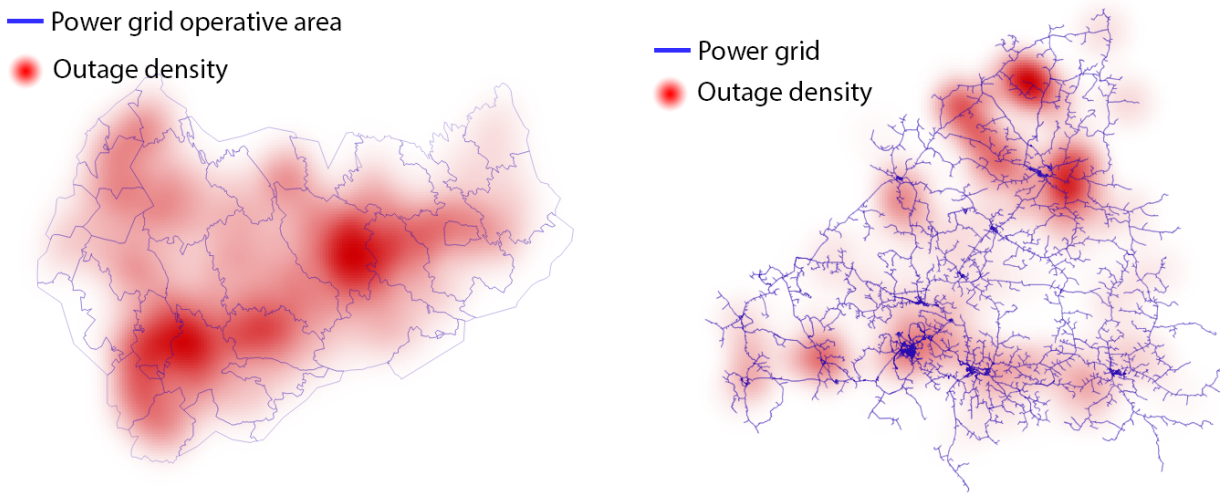
end for

Table 1. Extracted features. Features used in the final classification marked as bold.

Feature	Aggregation	Explanation
Speed	-	Object movement speed
Angle	-	Object movement angle
Area	-	Object size
Area difference	-	Object area difference to the previous time step
Week	-	Week of the year
Snowdepth	average, minimum, maximum	Snow depth
Total column water vapor	**average, minimum, maximum**	Total amount of water vapour
Temperature	**average, minimum, maximum**	2 meter air temperature
Snowfall	average, minimum, maximum, sum	Snowfall (meter of water equivalent)
Total cloud cover	average, minimum, maximum	Total cloud cover (0-1)
CAPE	**average**, minimum, **maximum**	Convective available potential energy (J/kg)
Precipitation kg/m2	average, minimum, maximum, sum	Precipitation amount (kg/m2)
Wind gust	**average**, minimum, **maximum, standard deviation**	Hourly maximum wind gust ($\mathrm{m\ s^{-1}}$)
Wind Speed	**average**, minimum, **maximum, standard deviation**	10 meter wind speed ($\mathrm{m\ s^{-1}}$)
Wind Direction	**average**, minimum, maximum, **standard deviation**	Wind direction (degrees))
Dewpoint	**average, minimum, maximum**	Dewpoint)
Mixed layer height	**average, minimum, maximum**	Boundary layer height
Pressure	average, **minimum**, maximum	Air pressure
Forest age	**average**, minimum, maximum, standard deviation	The age of the growing stock on a forest stand
Forest site fertility	**average**, minimum, maximum, standard deviation	Group of the forest by vegetation zones
Forest stand mean diameter	**average**, minimum, maximum, standard deviation	Forest stand mean mean diameter
Forest stand mean height	**average**, minimum, maximum, standard deviation	Forest stand mean height
Forest canopy cover	**average**, minimum, maximum, standard deviation	Forest canopy cover fraction (0-100%)
Outages	-	Number of occured outages
Customers	-	Number of affected customers
Tansformers	-	Number of transformers under the object
All customers	-	Number of customers under the object
Class	-	Assigned class

3.3 Defining classes

As shown in Figures 2a and 2b, the outages in the local dataset are concentrated heavily on 'hot-spots', assumingly, due to forest characteristics and network topology. The local dataset contains 24 542 storm objects and 5 837 outages connected to 2 363 storm objects. Thus 22 179 storm objects in the local dataset did not cause any outages. The local power outage data contain 16 191 outages, which can not be connected to any storm object. The national dataset contains 142 873 storm objects and 5 965 324 outages connected to 33 796 storm objects. 109 077 storm objects are not connected to any outages, and 175 110 outages can not be connected to any storm object.



(a) (b)

Figure 2. Spatial distribution of the outages between 2010 and 2018 visualised as a spatial heatmap. (a) JSE network (southern area) (b) Loiste network (northern area)

It should be noticed that the damage may occur anywhere in the power grid. Outages are, however, always reported as transformers without electricity. Typically one physical damage between the transformers causes several transformers to lose power. Power grid operators can often turn part of the transformers back to operation even before fixing the actual damage, which causes an unavoidable noise to the datasets.

Figure 3 represents the number of outages and storm objects in both local and national datasets. We can identify a large amount of 15 $\mathrm{m\,s^{-1}}$ storm objects in both sets, indicating that moderate wind without other influencing factors does not damage the transformers. When identifying storm objects with the contour of 20 and 25 $\mathrm{m\,s^{-1}}$, the number of objects reduces and starts to correlate more with a high number of outages, which supports views of previous studies showing the significance of stronger wind gusts to more severe storm damages. The method seems to identify also the most critical storm days by capturing several storm objects for those days. For instance, at the end of 2013, when the three major storms Eino, Oskari,

Seija (Valta et al., 2019) hit Finland, both datasets contain plenty of storm objects with the 20 $\mathrm{m\ s^{-1}}$ threshold. Nevertheless, our experiments indicated that employing 15 $\mathrm{m\ s^{-1}}$ storm objects yielded the best results. This is described more in Chapter 4.

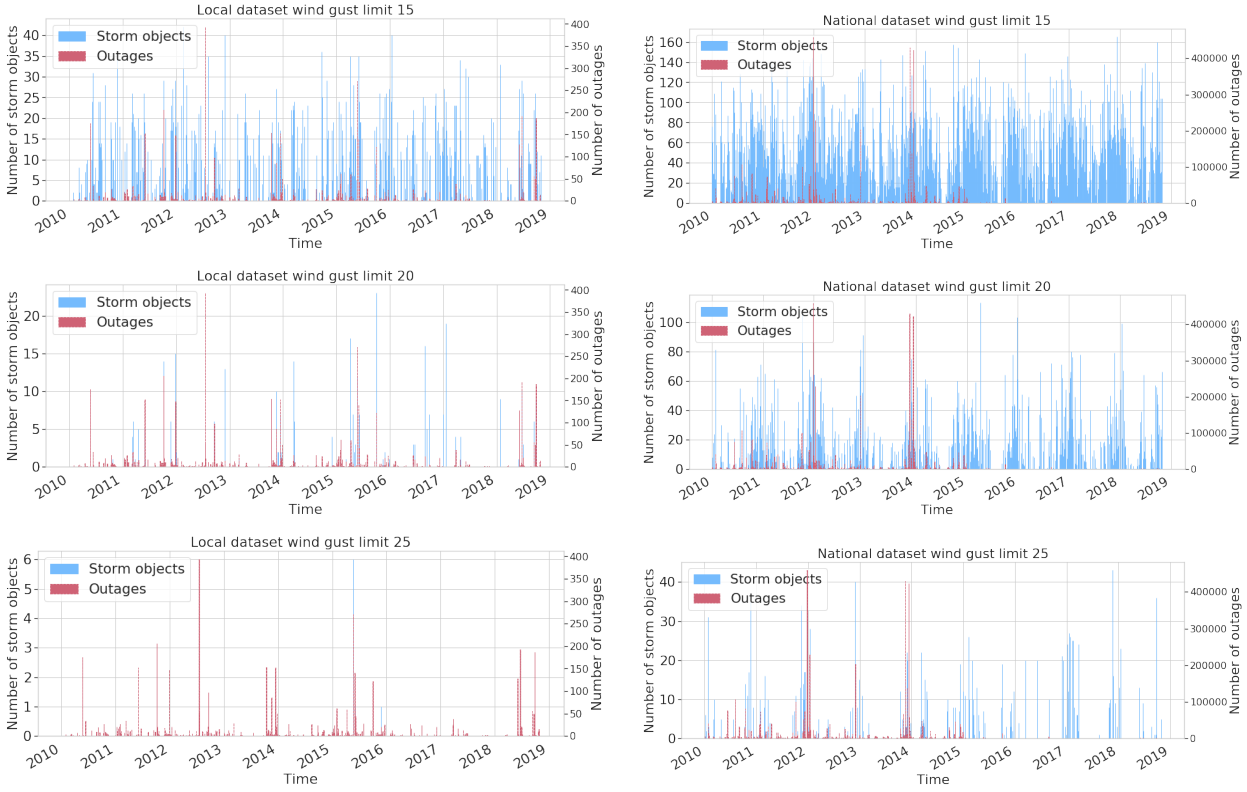
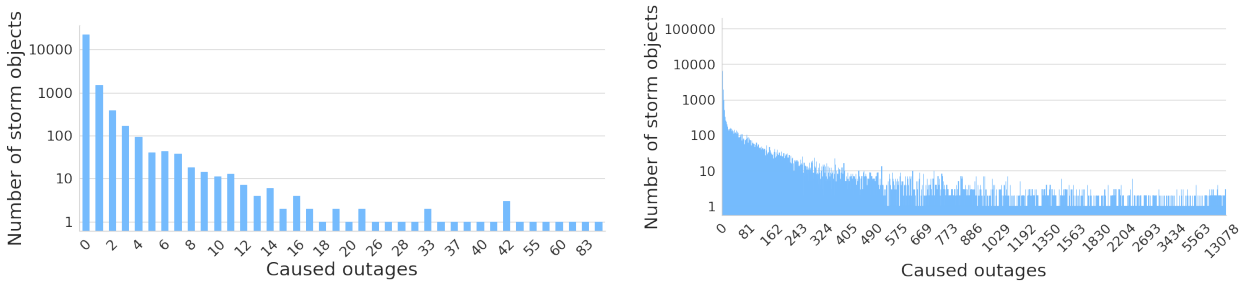


Figure 3. Storm object time series (15, 20 and 25 $\mathrm{m\ s^{-1}}$ contours) with occurred outages for local and national datasets.

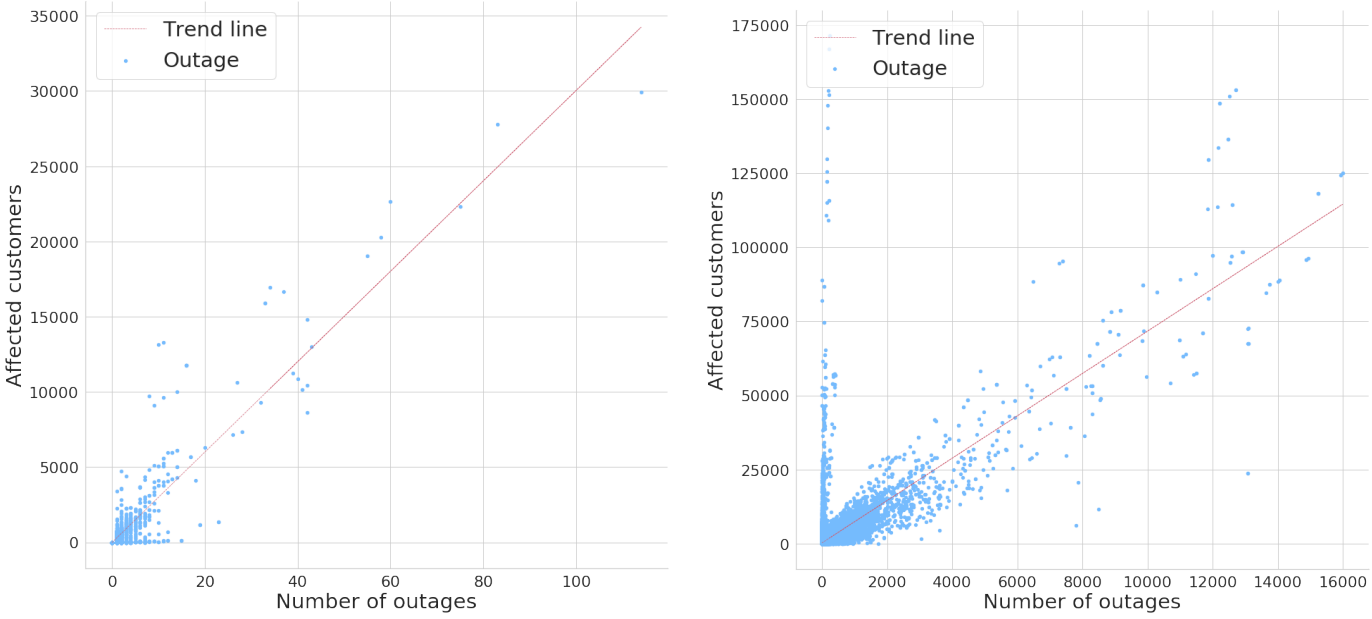
Figure 4 illustrates how much outages a single storm object typically produces. In the local dataset, most of the storm objects cause only a few outages. Only 65 storm objects, which are only 0.3 percent of the whole dataset, induced more than ten outages. On the other hand, in the national dataset where one storm object typically affects several different transformers, 17 587 storm objects have caused more than ten outages, representing 12 percent of the whole dataset. Figure 5 renders how many customers are typically affected by one outage. The figure contains all outages in both datasets, whether they are related to a storm or not. In the local dataset, usually 20-30 customers lose electricity in one outage. In the national dataset, only six customers usually lose electricity in one outage. We assume that this roots to different network topologies between the areas. Notably, in some rare cases, a much higher number of customers are affected. We assume that these cases occur typically in urban areas and are rare because the power network is mainly underground in these areas.

We use three classes designed together with power grid companies aiming at a simple "at glance" view for power grid operators. Class 0 represents no damage, class 1 low damage, and class 2 high damage. As the number of outages produced by a single storm object varies significantly in the local and national datasets, we decided to define separate limits for the local



(a) (b)

Figure 4. Number of storm objects per caused outages in (a) local dataset (b) national dataset.



(a) (b)

Figure 5. Relationship between number of outages and affected customers in (a) local dataset and (b) national dataset.

and the national datasets. The detailed limits are listed in Table 2. Class 1 is defined such that it represents roughly 80 percent of all cases with at least one outage. Class sizes are highly imbalanced as most of the storm objects do not cause any damage.

3.4 Classifying storm objects

We centered and normalized the data points by substracting the empirical mean and then dividing it by the empirical standard deviation. The hyperparameters were determined using random search 5-fold cross-validation (Bergstra and Bengio, 2012). To cope with the imbalanced class distribution, we generate artificial training samples using the synthetic minority over-sampling

Table 2. Class definitions

Class	Outage limit in local dataset	Local dataset size	Outage limit in national dataset	National dataset size
0	0	5 624	0	76 215
1	1-3	353	1- 140	14 417
2	≥ 4	181	≥ 141	3 085

technique SMOTE (Chawla et al., 2002). The SMOTE creates new training samples based on their $k = 5$ nearest neighbors following:

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i) \tag{1}$$

where x_i is an original class sample, x_{zi} is one of x_i's k nearest neighbor and λ is a random variable drawn uniformly from the interval $[0, 1]$. After augmentation, all classes have an equal number of samples, which reduces the tendency of classification methods to always predict the majority class.

Five different models were evaluated to classify storm objects. We omit the mathematical definitions but shortly discuss the characteristics of different models and describe the implementation details chosen in this work.

Random forest classification (RFC) is based on a random ensemble of decision trees and aggregating results from individual trees to the final estimate. Trees in the ensemble are constructed with four steps: 1) use bootstrapping to generate a random sample of the data, 2) randomly select a subset of features at each node, 3) determine the best split at the node using loss function, 4) grow the full tree (Breiman, 2001). RFC is good to cope with high-dimensional data. It has also been found to provide adequate performance with imbalanced data (Tervo et al., 2019; Brown and Mues, 2012) and is widely used with weather data (e.g. Karthick et al. (2020); Cerrai et al. (2019); Lagerquist et al. (2017)). The method is prone to overfit, which is why hyperparameter-tuning is very important. Hyperparameters used in this work are listed in Table 3. We use RFC with the Gini impurity loss function.

Table 3. Hyperparameters for the RFC

Parameter	Value
Number of trees in the forest	500
Max depth	unlimited
Minimum nr. of samples to split	2
Minimun nr of samples to leaf	1
Features to consider for split	$\sqrt{\text{num. of feat.}}$
Max nro of leaf nodes	unlimited

Support Vector Classifiers (SVC) construct a hyper-plane or classification function in a high-dimensional feature space and maximize a distance between training samples and the hyperplane. The hyper-planes may be constructed with nonlinear kernels such as gaussian radial basis function (RBF) (Shawe-Taylor et al., 2004) that often reform a nonlinear classification problem to a linear one. Operating in the high-dimensional feature space without additional computational complexity makes SVC an attractive choice to extract meaningful features from a high-dimensional dataset. A domain-specific expert knowledge can also be capitalized on the kernel design. On the other hand, finding the correct kernel is often a difficult task. Training SVC is a convex optimization problem meaning that it has no local minima. Depending on the kernel, a training process may, however, be a very memory-intensive process.

Suppose the SVM output is assumed to be the log odds of a positive sample. In that case, one can fit a parametric model to obtain the posterior probability function and thus get probabilities for samples to belong to the particular class (Platt et al., 1999). For more details, we request the reader to consult for example Chang and Lin (2011) and Platt et al. (1999).

We implement the SVC in two phases. First, we separate class 0 (no outages) and other samples employing SVC with radial basis function (RBF), defined in Equation 2. Second, we distinguish classes 1 and 2 using SVC with a dot-product kernel defined in Equation 3 (Williams and Rasmussen, 2006). The second phase is performed only for the samples predicted to cause outages in the first phase. The approach is similar to the often-used one-vs-one classification, where a binary classifier is fitted for each pair of classes. In our case different kernels were used for different pairs.

$$k_{RBF}(\mathbf{x}, \mathbf{x}') = \exp\big(-\gamma||\mathbf{x} - \mathbf{x}'||^2\big) \tag{2}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are two samples in the input space and γ is a kernel coefficient parameter.

$$k.(\mathbf{x}, \mathbf{x}') = \sigma_0 + x \cdot x' \tag{3}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are two samples in the input space and σ is a kernel inhomogenity parameter.

Gaussian Naive Bayes (GNB) (Chan et al., 1979) is a well-known and widely used method based on the Bayesian probability theory. The method assumes that all samples are independent and identically distributed (i.i.d), which does not naturally hold for the weather data. Despite the internal structure of the data, GNB is still used for weather data (e.g. Kossin and Sitkowski (2009); Cintineo et al. (2014); Karthick et al. (2020)) and worth investigating in this context. The classification rule in GNB is $\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y)$, where $P(y)$ is a frequency of class y and $P(x_i \mid y)$ is a likelihood of the ith feature assumed to be gaussian. Because of the naive i.i.d assumption, each likelihood can be estimated separately, which helps to cope with a curse of dimensionality and enable GNB to work relatively well with small datasets. On the other hand, estimating likelihoods can be done effectively and iteratively, enabling the GNB to scale to large datasets. As a downside, the simple method may lack expression power to perform well in a complex context.

Gaussian Processes (GP) (Rasmussen, 2003) is a non-parametric probabilistic method that interprets the observed data points as realizations of a Gaussian random process. GP is widely used for example in weather observation interpolation

kriging (Holdaway, 1996). GP is a very flexible and powerful but computationally expensive method, which tends to lose its power with high-dimensional data. GP hinges on a kernel function that encodes the covariance between different data points. As a kernel, we use a product of dot-product kernel (Equation 3) and pairwise kernel with laplacian distance (Rupp, 2015), defined in Equation 4. The kernel parameters were optimized on the training data by maximizing the log-marginal-likelihood.

$$k_{pairwise}(\mathbf{x}, \mathbf{x}') = \exp\big(-\gamma||\mathbf{x} - \mathbf{x}'||_1\big) \tag{4}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are two samples in the input space and γ is a kernel coefficient parameter.

Multilayer perceptrons (MLP) (Goodfellow et al., 2016) are the most basic form of artificial neural networks. Good results achieved by MLP in predicting storms (Ukkonen and Mäkelä, 2019), they are a natural choice to experiment in this work. Neural networks are very adaptive methods as they can learn a representation of the input at their hidden layers. Unlike GNB, they do not make any assumptions about the distribution of the data. As a downside, MLP requires large amounts of data, and the training process is computing-intensive. They also have a large number of hyperparameters to be optimized, including the correct network topology.

We searched the correct model parameters and network topology for local and national datasets by running multiple iterations of random search 5-fold cross-validation to obtain the best possible micro average of F1-score (defined in Chapter 4) employing Talos library (Autonomio, 2020). The final setup composes of Nadam optimizer (Dozat, 2016), random normal initializer, and relu activation function for hidden layers. Binary cross-entropy was used as a loss function. Optimal network topology varied in different datasets: For the local dataset, the best results were obtained with a network containing three hidden layers with 75, 145, and 35 neurons. For the national dataset, the best results were obtained with a network containing three hidden layers with 75, 195, and 300 neurons. During the optimization process, the results varied between different setups from 0.6 to 0.95 in terms of F1-score.

4 Results

We used two different methods for splitting the data into training and test set. The first method is to use 25 percent of randomly picked samples in the test set. The second method is to construct a test set from a one-year continuous time range (2010-2011). Both approaches have their advantages. Continuous time range ensures that the model has not seen any autocorrelated samples caused by an internal structure of the weather data in the training phase (Roberts et al., 2017). However, having only nine years of data from a relatively small geographical area, the continuous test set cannot contain many storms as most of the data needs to be reserved for the training process. Thus, the test set may only contain a single type of storms to which the model may work especially well or bad. Picking the test set randomly minimizes this risk and provides more insight into the model performance.

We evaluate the models with a weighted average of precision and recall, and both weighted and macro average of F1-score. Precision (Equation 5) reports how many samples are correctly predicted to belong to a class. Recall (Equation 6) tells how many samples belonging to a class are found in the prediction. F1-score (Equations 7 and 8) calculates a harmonic mean of

precision and recall. Finally, as the datasets are extremely imbalanced, we calculate a weighted average of the metrics utilizing a number of samples in each class and a macro average of F1-score using an average of F1-score of each class. A model with a higher macro average of F1-score performs better with small classes. The selected metrics do not take a distance between predicted and true class into account. It is naturally worse to predict, for example, class 0 (no damage) in the case of true class 2 (high damage) than in the case of true class 1 (low damage). We decided, however, to use metrics that measure the method performance properly with imbalanced classes.

$$Precision = \frac{1}{\sum_{c \in \mathcal{C}} |\hat{y}_c|} \sum_{c \in \mathcal{C}} \left(|\hat{y}_c| \frac{tp}{tp + fp} \right) \tag{5}$$

where $\mathcal{C}$ represents the set of classes, $\hat{y}$ predicted the class, tp true positives, and fp false positives.

$$Recall = \frac{1}{\sum_{c \in \mathcal{C}} |\hat{y}_c|} \sum_{c \in \mathcal{C}} \left(|\hat{y}_c| \frac{tp}{tp + fn} \right) \tag{6}$$

where $\mathcal{C}$ represents the set of classes, $\hat{y}$ predicted the class, tp true positives, and fn false negatives.

$$F1_{weighted} = \frac{1}{\sum_{c \in \mathcal{C}} |\hat{y}_c|} \sum_{c \in \mathcal{C}} \left(|\hat{y}_c| \frac{Precision_c \times Recall_c}{Precision_c + Recall_c} \right) \tag{7}$$

where $\mathcal{C}$ represents the set of classes, $\hat{y}$ predicted the class, Precision defined in Equation 5, and Recall defined in Equation 6.

$$F1_{macro} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left(\frac{Precision_c \times Recall_c}{Precision_c + Recall_c} \right) \tag{8}$$

where $\mathcal{C}$ represents the set of classes, Precision defined in Equation 5, and Recall defined in Equation 6.

Tables 4 and 5 divulge the results for each models using the local and national dataset respectively. Models trained with the local dataset can reach the better-weighted F1-score, while the best models trained with the national dataset provide a significantly better macro average of F1-score. The national dataset contains many more samples in classes 1 and 2, which enables models to learn the classes better and thus enhance the macro average of the F1-score. Whether the test set is randomly chosen or continuous does not seem to make a large difference in most cases. The only affected model is the RFC having contradictory better results trained with the continuous test set from the local dataset and the random test set from the national dataset. Assumingly, this squeal more about the unstable performance of RFC than the relevance of the dataset split method.

The confusion matrices are depicted in Figure 6. RFC provides the best results in terms of the selected metrics. However, closer exploration reveals that this performance is largely due to the best performance in predicting class 0, which is the largest class. SVC results are one of the most balanced ones being the best only in the local dataset with a random test set but yielding good stable results in all cases. The confusion matrix, shown in Figure 6b, displays that it is not the best model to predict class

Table 4. Results for each models trained with the local dataset obtained from two local power grid companies (defined in Chapter 3.3)

Model	Split method	Precision	Recall	Weighted F1-score		Macro AVG F1-score	
		test	test	train	test	train	test
Random Forest Classifier (RFC)	Random	0.82	0.76	0.93	0.79	0.93	0.40
	Continuous	0.88	**0.91**	0.93	**0.89**	0.93	**0.48**
Support Vector Classifier (SVC)	Random	0.85	0.73	0.78	0.78	0.78	**0.44**
	Continuous	0.87	0.72	0.77	0.78	0.77	0.42
Gaussian Naive Bayes (GNB)	Random	**0.87**	0.61	0.59	0.70	0.59	0.42
	Continuous	**0.89**	0.59	0.59	0.69	0.59	0.40
Gaussian Processes (GP)	Random	0.84	0.70	1.0	0.76	1.0	0.43
	Continuous	0.85	0.67	0.94	0.74	0.94	0.41
Multilayer perceptor (MLP)	Random	0.82	**0.81**	0.98	**0.80**	0.91	0.41
	Continuous	0.81	0.79	0.97	0.80	0.91	0.41

Table 5. Results for each models trained with the national dataset covering whole Finland (defined in Chapter 3.3)

Model	test set split method	Precision	Recall	Weighted F1-score		Macro AVG F1-score	
		test	test	train	test	train	test
Random Forest Classifier (RFC)	Random	**0.83**	**0.84**	1.0	**0.83**	1.0	**0.62**
	Continuous	**0.77**	**0.81**	1.0	**0.78**	1.0	0.40
Support Vector Classifier (SVC)	Random	0.81	0.61	0.68	0.68	0.68	0.60
	Continuous	0.62	0.60	0.60	0.60	0.60	0.60
Gaussian Naive Bayes (GNB)	Random	0.75	0.60	0.66	0.66	0.45	0.39
	Continuous	**0.77**	0.60	0.45	0.66	0.45	0.40
Gaussian Processes (GP)	Random	0.57	0.56	0.71	0.55	0.71	0.55
	Continuous	0.67	0.65	0.94	0.65	0.94	**0.61**
Multilayer perceptor (MLP)	Random	0.79	0.75	0.94	0.77	0.90	0.52
	Continuous	0.76	0.78	0.93	0.78	0.85	0.40

0, but only a little share of true class 2 cases and the smallest share of true class 1 cases are predicted as class 0. That is to say,
310 SVC misses the smallest number of destructive storms, although it confuses in the amount of caused damage.

GP is another strong option that performs even better with class 0 while still providing good performance with class 2. A significant connecting aspect between GP and SVC is an almost identical kernel. Based on these experiments, RBF and pairwise kernels separate harmless and harmful samples from each other while dot-product kernel separates the classes 1 and 2 even better than exponential functions. We select GP for further analysis in this paper since it provides the best performance in class 2.

Using the 15 $\mathrm{m\,s^{-1}}$ threshold for detecting storm objects yields clearly better results than the 20 $\mathrm{m\,s^{-1}}$ threshold. For example, SVC trained with the national dataset using the 20 $\mathrm{m\,s^{-1}}$ threshold and randomly chosen test set provide only 0.48 macro average of F1-score being 12 percentage points below corresponding model using the 15 $\mathrm{m\,s^{-1}}$ threshold. The 15 $\mathrm{m\,s^{-1}}$ threshold have two major advantages compared to the 20 $\mathrm{m\,s^{-1}}$. First, it provides a significantly larger dataset and second, in contrast to the 20 $\mathrm{m\,s^{-1}}$ threshold, it is able to catch virtually all extratropical storms causing outages.

4.1 Feature importances in the model performance

The relevance of the individual predictive features can be explored by using the permutation test, as done by Breiman (2001). First, the baseline score of the fitted model is calculated using the test set. Then each feature is randomly permuted, and the difference in the scoring function is calculated. The random permutation is repeated 30 times for each parameter, and the average of the results is used. The procedure offers information on how important the feature is to obtain good results. It should be mentioned that highly correlated features may get low importance as other features work as a proxy to the permuted feature. However, using completely independent features is not possible in weather data since weather parameters are often dependent on each other, and eliminating even the most apparent pairs from the used features impaired the results in our experiments.

We used the macro average of F1 defined in Equation 8 as a scoring function and the randomly selected test set from the national data. The relevance is shown in Figure 7. Most features show at least little relevance for the results. The first twelve features are significantly more relevant than the rest. The most important features contain at least one representative of all meteorological parameters used in the training. In other words, all employed meteorological parameters are important for the prediction, while different aggregations are contributing to the "fine-tuning" of the model.

As Figure 7 shows, the most significant parameter regarding our model performance is the average wind speed. Numerous studies support our result of wind being the most important damaging factor (Virot et al., 2016; Valta et al., 2019; Jokinen et al., 2015). The studies are, however, highlighting the importance of maximum wind gusts instead of the average wind. Surprisingly, in our analysis, the wind gust speed does not belong to the most critical parameters. Instead, maximum mixed layer height, related to the wind gustiness, contributes crucially to the model performance. The dependencies between predictive features might be one reason for some parameters to have a lower rank in the results.

The stand mean diameter and height are the most important features regarding the forest parameters, which corresponds to our expectations. Previous studies also state these features to influence the wind damage in forests (Pellikka and Järvenpää, 2003) and hence indirectly electricity grids. As Pellikka and Järvenpää (2003) and Suvanto et al. (2016) discuss, also the age of the forest has an impact on storm damages. However, in the feature importance test, forest age does not seem to contribute significantly to the prediction outcome.

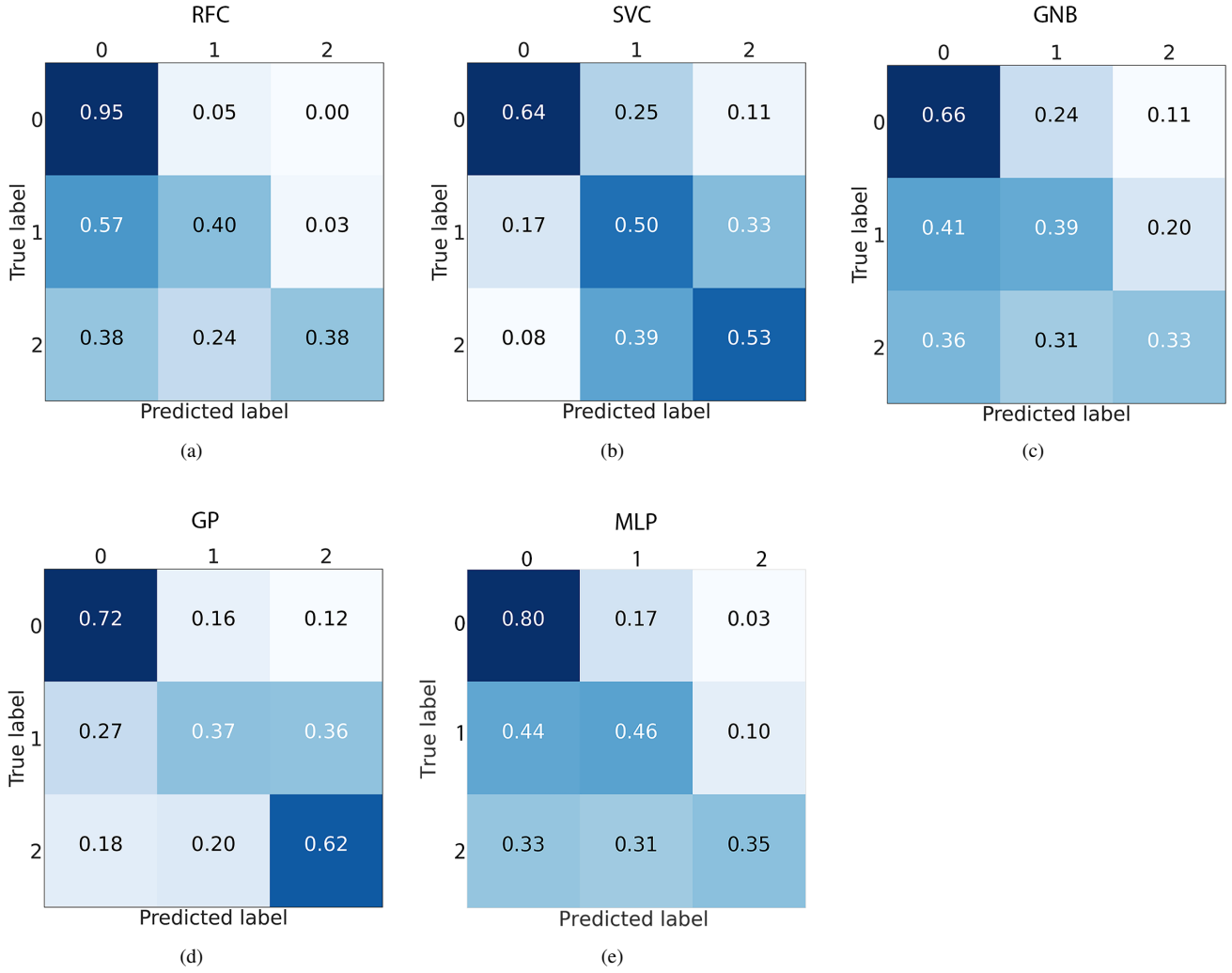


Figure 6. Confusion matrices produced using the randomly selected national dataset and (a) RFC (b) SVC (c) GNB (d) GP (e) MLP. Each cell of the confusion matrices represents a share of predictions having a corresponding combination of predicted and true class. For example, the middle right cell tells the share of samples belonging to class 1 but predicted to have class 2.

345 The most important object feature is the size of the object. Object movement speed and direction did not contribute strongly to the results. However, previous studies indicate that besides the size of the impacted area, the duration of strong winds – i.e., the propagation speed of the system – influences also the amount of damage (Lamb and Knud, 1991).

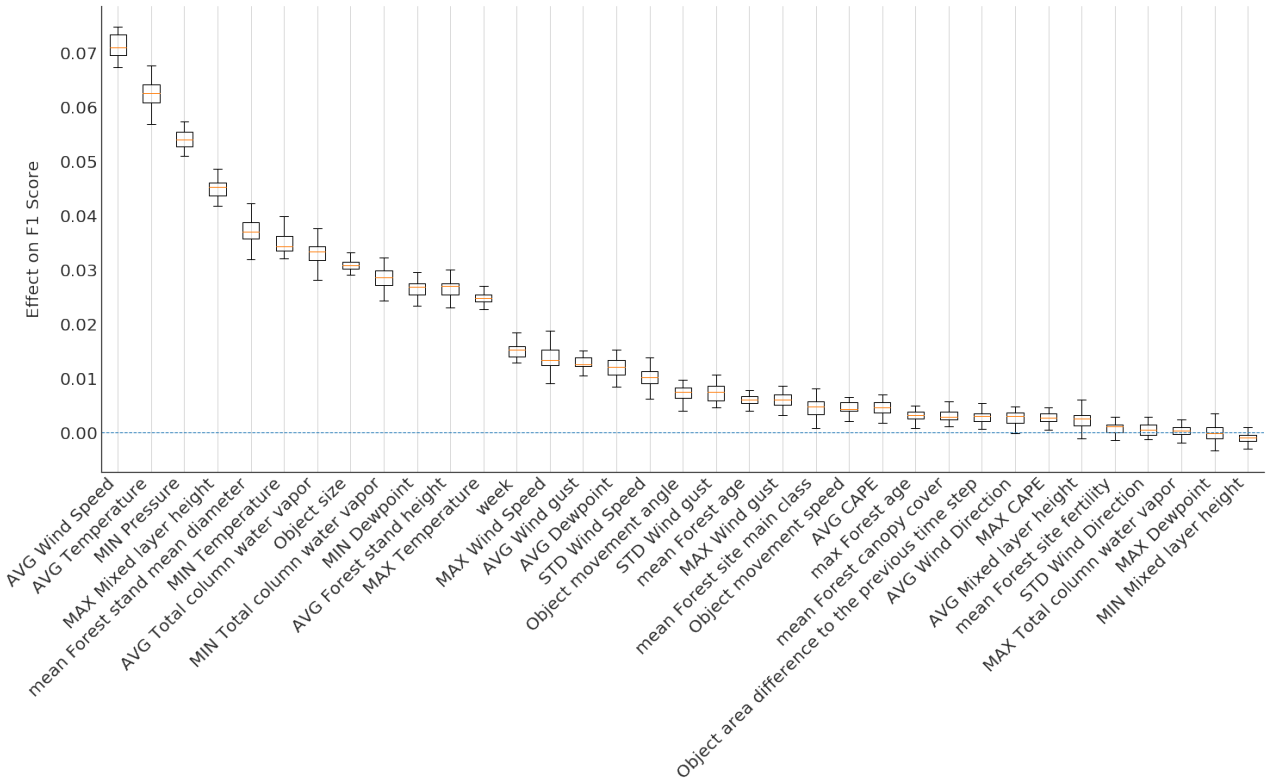


Figure 7. Permutation feature importance using the GP classification method trained with the randomly selected national dataset. The higher effect on the F1 score is (y-axis), the bigger is the significance.

4.2 Case Examples

We illustrate the prediction produced using GP classification method with the three most interesting examples of well-known storms in Figure 8a. We chose the cases among a number of test cases to illustrate the strengths and weaknesses of the method. The examples are chosen from the randomly picked test set, which was not used to train the model. Because of the random sample, we cannot represent the entire prediction of individual storms, only individually picked time steps. In two of the example cases, the model performs well (storms Tapani and Pauliina) and in one (storm Rauli) with less accurate prediction results.

4.2.1 Event 1: Extratropical Storm Tapani (26 December 2011)

The first example is one of the most known extratropical storms in Finland. Storm Tapani, also known as Cyclone Dagmar (Kufeoglu and Lehtonen, 2015), was a rare winter storm, causing broad and long-lasting electricity interruptions. Extreme wind gusts of over 30 $\mathrm{m\ s^{-1}}$ caused widespread damage, especially in the southern and western parts of the country. Approximately 570 000 households were left without electricity, causing 30 million euros repair costs and 80 million euros of monetary

compensation for electricity distribution companies to their customers (Hanninen and Naukkarinen, 2012). Exceptionally warm December and the Boxing day being the warmest in 50 years (Finnish Meteorological Institute, 2011) resulted in wet and unfrozen soil. Thus, the trees were poorly anchored and exposed to significant storm damage.

Figure 8a represents the outage prediction (raster-covered areas) and the actual, true classes (numbers) based on the damage data at 15:00 UTC, 26 December 2011. Wide areas in central and western parts of Finland are predicted to have *high* (class 2) damages. The predicted class is in line with the true class. Also, the damage areas of the storm correlate with the wind gust observations of the Finnish Meteorological Institute. The strongest gusts occurred in western (15-27 $\mathrm{m\,s^{-1}}$) and southern (18-28 $\mathrm{m\,s^{-1}}$) Finland and north-western part of Lapland (13-31$\mathrm{m\,s^{-1}}$) (Finnish Meteorological Institute, 2020). In the rest of Finland, the maximum wind gusts remained between 10-15 $\mathrm{m\,s^{-1}}$, and therefore the damages were minor. Overall, the model predicted the damages accurately in this particular example.

4.2.2 Event 2: Extratropical Storm Rauli (27 August 2016)

Extratropical storm Rauli was an exceptionally strong summer storm, especially regarding the impacts. It caused severe damages to the power grid in the western and middle parts of Finland for various reasons. The trees were carrying leaves, the soil was wet after a rainy August, the strong wind areas of Rauli were widely spread, and the solar radiation was intensifying the wind gusts during the afternoon (Finnish Meteorological Institute, 2016). Rauli was impacting especially the middle and southern parts of Finland, which are also the most densely populated areas. The power outages were increasing rapidly in the middle part of Finland, starting at midday and reaching the highest values, 200 000 households without electricity (Ilta-Sanomat, 2016), around 5 pm. The winds were blowing exceptionally long, nearly 24 hours. The typical duration of summer storms is between 6-12 hours.

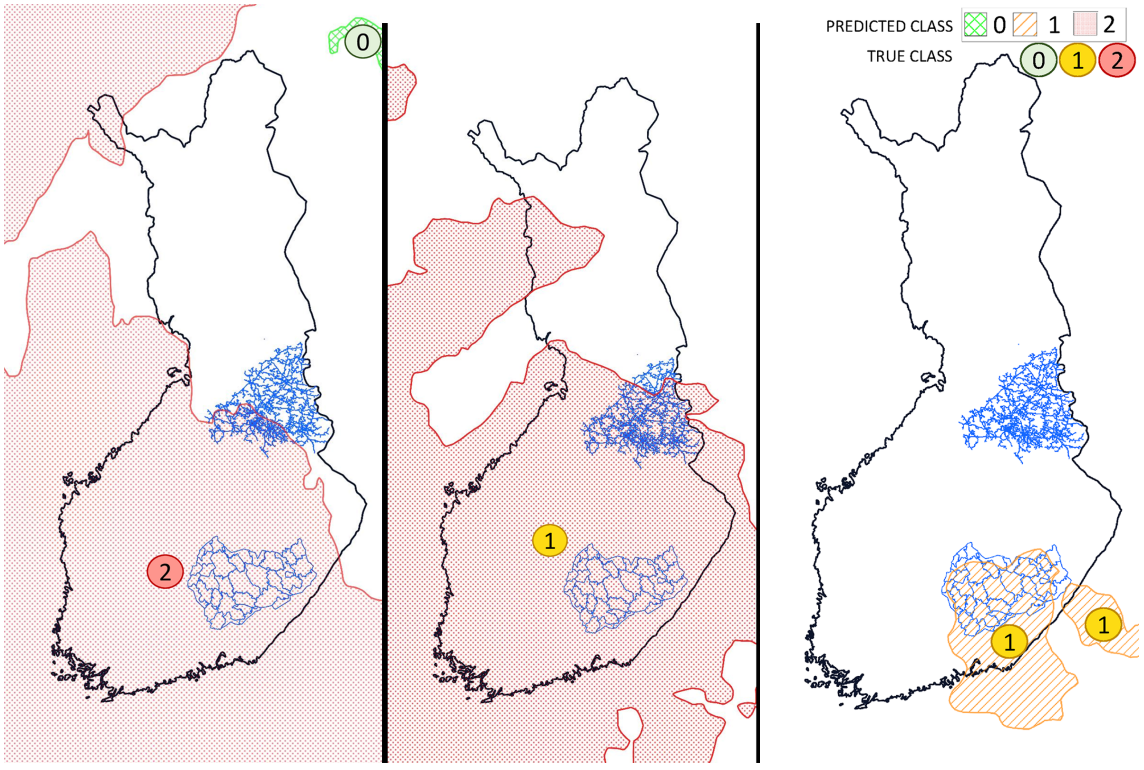
Figure 8b shows the predicted outages and true classes at 12:00 UTC, 27 August 2016. In this particular time step, the model is over-predicting the class, however, the predicted outage area seems to correlate with the wind gust maximums of that afternoon. The strongest wind gusts were measured in the southern and middle parts of the country, maximum gusts reaching on land stations up to 24,9 $\mathrm{m\,s^{-1}}$ (Klemettilä, Vaasa and Maaninka, Pohjois-Savo) and on wide areas up to 20 $\mathrm{m\,s^{-1}}$ apart from the northern part of Finland.

4.2.3 Event 3: Extratropical storm Pauliina (22 June 2018)

The last example is a strong extratropical storm, called Pauliina (Finnish Meteorological Institute, 2018) that caused numerous power outages in Finland. The most significant part of the power outages happened in the network of power grid company JSE included in the local dataset. The highest peak in the damages was reached between 6 and 8 pm with over 28 000 households without electricity. The strongest wind gust on land reached 22,7 $\mathrm{m\,s^{-1}}$ in Helsinki, Kumpula observation station, and the inland gusts were widely between 15-20 $\mathrm{m\,s^{-1}}$ (Finnish Meteorological Institute, 2020; Finnish Meteorological Institute, Twitter). The strong wind gusts continued until the dawn of the 23rd of June.

Figure 8c presents the predicted and true damage classes at 01:00, UTC, 22 June 2018. We chose extratropical storm Pauliina as an example storm for two reasons: 1) Pauliina represents a *low damage* class 2) Pauliina represents a rare, summer-season

extratropical storm. Figure 8c shows the predicted and true classes correlating. While weather warnings were issued to large areas in southern and middle parts of Finland, (myrskyvaroitus.com, 2018) predicted and true damage to the power grid occurred in a relatively small geographical area.



(a) Tapani, 26 December 2011 11:00 (b) Rauli, 27 August 2016 10:00 (c) Pauliina, 22 June 2018 01:00

Figure 8. Selected examples (a) Extratropical storm Tapani (b) Extratropical storm Rauli (c) Extratropical storm Pauliina, produced by employing the SVC model trained with the national dataset. The storm objects are colored based on the predicted class while the true class is stated as a colored number over the object. The time is represented as UTC time.

5 Discussion and conclusions

This paper introduces a novel method to predict the damage potential of extratropical storms to power grids. The method consists of identifying storm objects by contouring surface wind gust fields with the 15 $\mathrm{m\,s^{-1}}$ threshold along with pressure objects with a 1000 hPa threshold, tracking the objects, and then classifying them into three classes based on their damage potential to the power grid. For the classification task, we evaluated five different machine learning methods, all employing in a total of 35 predictive features and trained with eight years of power outage data from Finland.

Both Gaussian Processes and Support Vector Classifiers provided good results. The model recognizes harmful storm objects well and can distinguish extremely harmful objects among others adequately. While the results still leave a lot to improve, the developed model can be already used to support decisions in power grid companies. In some cases, the model is able to provide a more specific and geospatially accurate prediction of potential damage to the power grid than, for example, weather warning. The evaluation was, however, based on the ERA5 reanalysis data. Using the method in an operational setting would require weather prediction data, which introduces additional uncertainty to the outage prediction.

The presented object-based approach has both advantages and disadvantages. Extracting storm objects in advance preprocesses the data for machine-learning techniques, such as RFC, which do not perform feature learning. It enables machine-learning methods to focus only on the relevant parts of the data. Methods not containing feature learning, such as RFC and logistic regression, have been found to outperform neural networks for forest (Hart et al., 2019) and weather data (Tervo et al., 2019). It also leads to significantly faster training times. Processing objects instead of the grid makes it also easier to track and use object attributes such as age, speed, and movement. Moreover, objects are easy to visualize, and user interfaces may be enriched with related actions such as tracking and alarms.

On the other hand, storm objects use only aggregated attributes, which may decrease the classification accuracy when predictive features vary significantly under the storm object area. Several machine-learning methods, i.e., deep neural networks, could be trained to employ those local features to gain better accuracy. Such methods could also utilize three-dimensional data.

Extracting storm objects requires a fixed threshold of wind gust and pressure, which may vary depending on the characteristics of geospatial locations. Nevertheless, the previous studies indicate the critical threshold to wind gust speed to be the same for almost entire geospatial domain of this work (Gardiner et al., 2013). Moreover, the correct threshold may vary depending on the data source. When extending the geospatial domain or changing the data source, this might become a more serious issue, and different thresholds might be needed. One possibility to determine the optimal threshold might be to use specific quantiles of the parameter values, but this would need further investigation.

The work opens several possible avenues for further studies. It would be interesting to compare the current solution with a grid-based approach and deep neural networks. Including data on soil moisture, soil temperature, and leaf index would most likely enhance the results, if available with sufficient spatial and temporal resolution, since they would provide critical information about the environmental conditions. Different thresholds could be investigated as well, especially for pressure objects where lower thresholds might yield better results. By design, applying the method to other regions is possible, but it is subject to the availability of power outage records, forest inventory, impact and meteorological data. For the classification task, carefully designed Bayesian networks could provide good results as well. Especially in the randomly selected test set, data may be autocorrelated, which may lead to unrealistically good results. We have addressed this issue by also using a continuous time series (from 2010 to 2011) for the test set. The evaluation could also be extended with a leave-one-day-out or leave-one-week-out method where for each week one day or for each month one week is hold out for validation purposes.

End users, especially expert users like duty forecasters, might benefit from the uncertainty information originating as the probabilistic prediction of the classification model. However, the presentation of such information should be very carefully chosen not to mislead non-expert users for overconfidence.

Experiments in this study were conducted with ERA5 reanalysis and additional forest data. As the method employs common features existing also in various other datasets, data provided by other vendors could be used as well. By employing weather forecasts as input, this method could be used as a base for a decision support tool and as a part of an existing early warning system for both duty forecasters of national hydro-meteorological centers and operators of electricity transmission companies.

Code and data availability. The source code is available in the repositories https://github.com/fmidev/sasse-era5-smartmet-grid and https://github.com/fmidev/sasse-polygon-process. ERA5 data may be downloaded from the Copernicus Climate Data Store: https://cds.climate.copernicus.eu. Forest inventory may be downloaded from LUKE open data service: http://kartta.luke.fi/index-en.html. The power outage data is propriety data which the authors have no property rights to distribute.

Appendix A: Gaussian distribution fitted to the storm object features

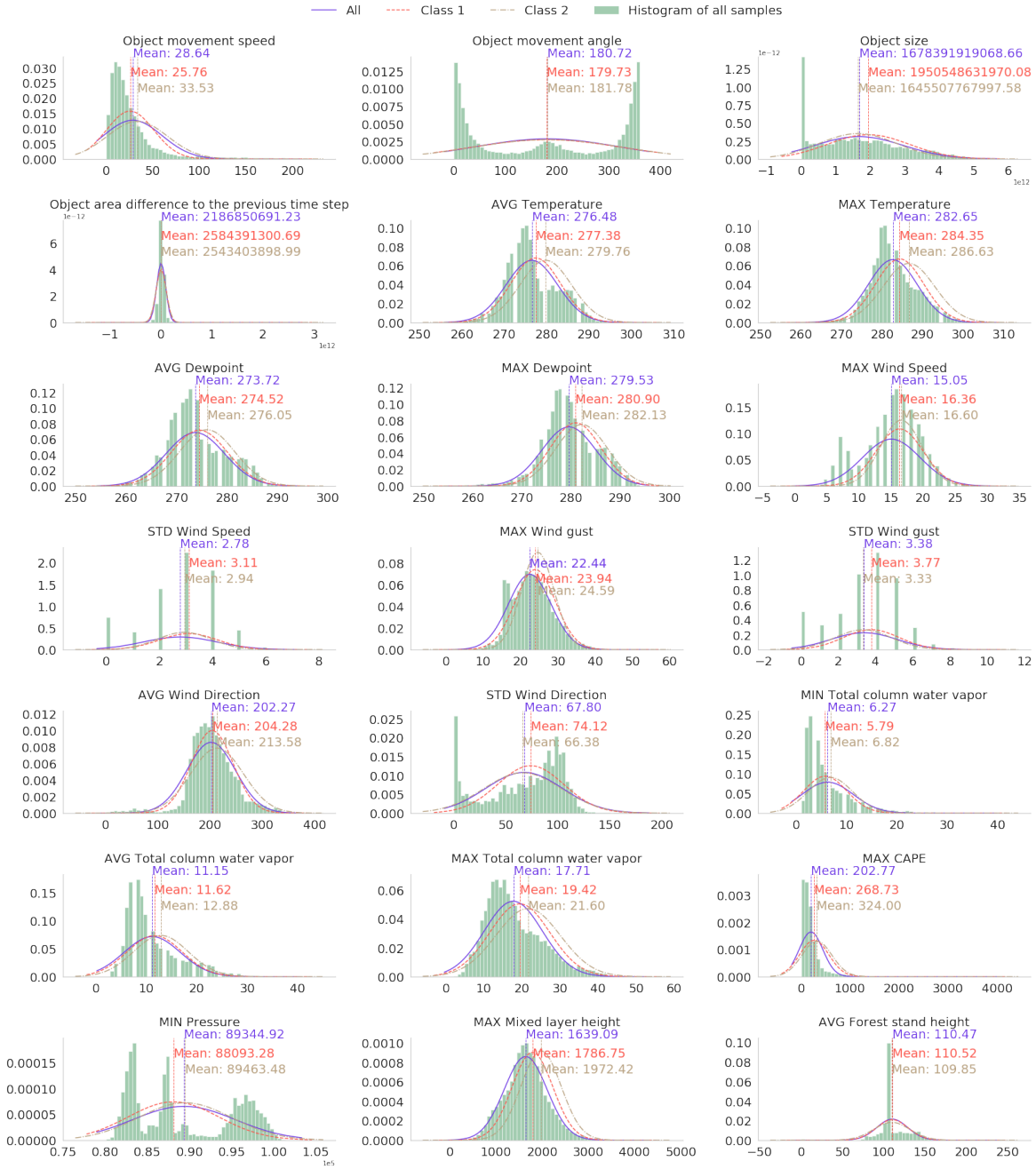


Figure A1. Histogram of and fitted Gaussian distribution of selected predictive parameters in **the local dataset**. The Gaussian distribution is fitted separately to all samples and samples with little outages and many outages (classes 1 and 2 specified in Section 3.3).

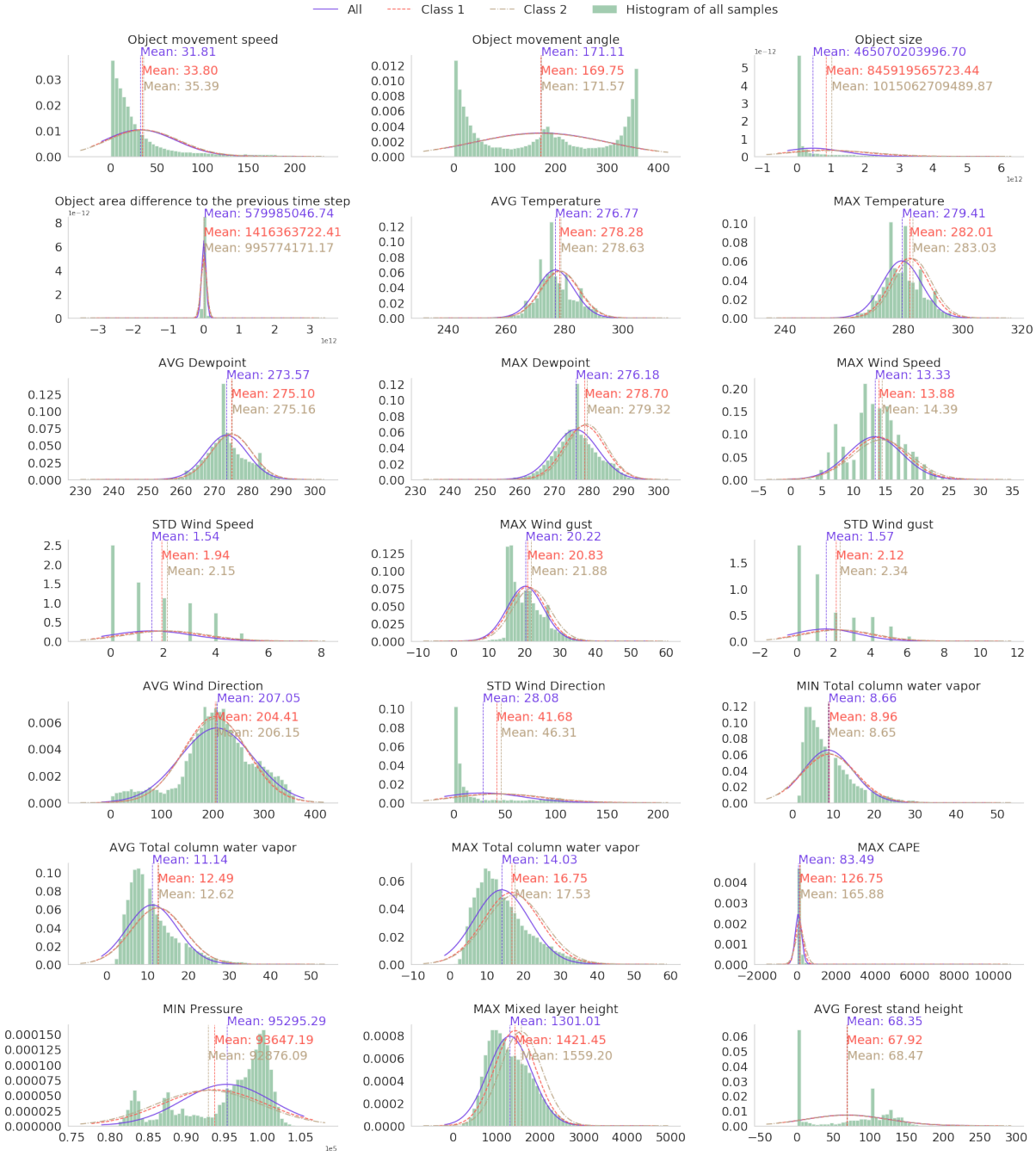


Figure A2. Histogram of and fitted Gaussian distribution of selected predictive parameters in **the national dataset**. The Gaussian distribution is fitted separately to all samples and samples with little outages and many outages (classes 1 and 2 specified in Section 3.3).

Author contributions. RT conceptualized, designed, and developed the method. IL contributed with meteorological expertise, such as selecting used data and meteorological features along with correct thresholds. IL also helped in analyzing the performance. AM provided supervision from a meteorological perspective and AJ from a machine learning perspective. All contributed in presenting the results.

Competing interests. The authors declare that they have no conflict of interests.

Acknowledgements. The authors express their gratitude to Järvi-Suomen Energia, Loiste Sähköverkko, and Imatran Seudun Sähkönsiirto for sharing data and their experience.

References

Allen, M., Fernandez, S., Omitaomu, O., and Walker, K.: Application of hybrid geo-spatially granular fragility curves to improve power outage predictions, Journal of Geography & Natural Disasters, 4, 1–6, https://doi.org/10.4172/2167-0587.1000127, 2014.

Autonomio: Talos (software), http://github.com/autonomio/talos., 2020.

Barredo, J.: No upward trend in normalised windstorm losses in Europe: 1970-2008, Natural Hazards and Earth System Sciences, 10, https://doi.org/10.5194/nhess-10-97-2010, 2010.

Bergstra, J. and Bengio, Y.: Random search for hyper-parameter optimization, Journal of Machine Learning Research, 13, 281–305, 2012.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Brown, I. and Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets, Expert Systems with Applications, 39, 3446–3453, 2012.

Campbell, R. J. and Lowry, S.: Weather-related power outages and electric system resiliency, Tech. rep., Congressional Research Service, Library of Congress Washington, DC, 2012.

Cerrai, D., Wanik, D. W., Bhuiyan, M. A. E., Zhang, X., Yang, J., Frediani, M. E., and Anagnostou, E. N.: Predicting storm outages through new representations of weather and vegetation, IEEE Access, 7, 29 639–29 654, 2019.

Chan, T., Golub, G., and LeVeque, R.: Updating formulae and a pairwise algorithm for variances computing sample, in: COMPSTAT, p. 30, Springer Science & Business Media, 1979.

Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, ACM transactions on intelligent systems and technology (TIST), 2, 27, 2011.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, 321–357, 2002.

Chen, P.-C. and Kezunovic, M.: Fuzzy logic approach to predictive risk analysis in distribution outage management, IEEE Transactions on Smart Grid, 7, 2827–2836, 2016.

Chen, S. H., Jakeman, A. J., and Norton, J. P.: Artificial intelligence techniques: an introduction to their use for modelling environmental systems, Mathematics and computers in simulation, 78, 379–400, 2008.

Cintineo, J. L., Pavolonis, M. J., Sieglaff, J. M., and Lindsey, D. T.: An empirical model for assessing the severe weather potential of developing convection, Weather and Forecasting, 29, 639–653, 2014.

Csilléry, K., Kunstler, G., Courbaud, B., Allard, D., Lassegues, P., Haslinger, K., and Gardiner, B.: Coupled effects of wind-storms and drought on tree mortality across 115 forest stands from the Western Alps and the Jura mountains, Global change biology, 23, https://doi.org/10.1111/gcb.13773, 2017.

Donat, M. G., Leckebusch, G. C., Wild, S., and Ulbrich, U.: Future changes in European winter storm losses and extreme wind speeds inferred from GCM and RCM multi-model simulations, Natural Hazards and Earth System Sciences, 11, 1351–1370, https://doi.org/10.5194/nhess-11-1351-2011, https://www.nat-hazards-earth-syst-sci.net/11/1351/2011/, 2011.

Dozat, T.: Incorporating nesterov momentum into adam, 2016.

Eskandarpour, R. and Khodaei, A.: Machine Learning Based Power Grid Outage Prediction in Response to Extreme Events, IEEE Transactions on Power Systems, 32, 3315–3316, 2017.

European Centre for Medium-Range Weather Forecasts: ERA5 Reanalysis, https://doi.org/10.5065/D6X34W69, 2017.

European Centre for Medium-Range Weather Forecasts: ERA5: large 10m winds, https://confluence.ecmwf.int/display/CKB/ERA5:+large+10m+winds, 2019.

Finnish Energy, E. O.: Energy distribution interruptions 2010-2018 - Finnish Energy, https://energia.fi/julkaisut/materiaalipankki/sahkon_keskeytystilastot_2010-2018.html#material-view, 2010-2018.

Finnish Energy, E. O.: Energy distribution interruptions 2011 - Finnish Energy, Tech. rep., Finnish Energy ry, https://energia.fi/files/605/Keskeytystilasto_2011.pdf, 2011.

Finnish Energy, E. O.: Energy distribution interruptions 2013 - Finnish Energy, Tech. rep., Finnish Energy ry, https://energia.fi/files/605/Keskeytystilasto_2013.pdf, 2013.

Finnish Meteorological Institute: Climate Statistics Report (December 2011), pp. 1–24, 2011.

Finnish Meteorological Institute: Rauli nousi myrskyjen raskaaseen sarjaan | Ilmastokatsaus – Ilmatieteen laitos, Ilmastokatsaus, http://www.ilmastokatsaus.fi/2016/08/29/rauli-nousi-myrskyjen-raskaaseen-sarjaan/, 2016.

Finnish Meteorological Institute: Kesäkuun 2018 kuukausikatsaus | Ilmastokatsaus – Ilmatieteen laitos, http://www.ilmastokatsaus.fi/2018/07/04/kesakuun-2018-kuukausikatsaus/, 2018.

Finnish Meteorological Institute: Open data, observation download service, https://en.ilmatieteenlaitos.fi/download-observations, 2020.

Finnish Meteorological Institute, Twitter: Ilmatieteen laitos Twitterissä: "#Juhannus'aaton-#myrsky #Pauliina lukuina klo 15.30: kovin puuska 32,5 m/s ja 10min keskituuli 26,5 m/s (Kaskinen Sälgrund), maa-alueilla Helsinki Kumpula puuska 22,7 m/s. Vaasa Klemettilä 24h sademäärä 41 mm." / Twitter, https://twitter.com/meteorologit/status/1010144460703969280, 2020.

Gardiner, B., Blennow, K., Carnus, J., Fleischner, P., Ingemarson, F., Landmann, G., Lindner, M., Marzano, M., Nicoll, B., Orazio, C., Peyron, J., Reviron, M., Schelhaas, M., Schuck, A., Spielmann, M., and Usbeck, T.: Destructive storms in European Forests: Past and Forthcoming Impacts. Final report to European Commission - DG Environment,, European Forest Institute, 2010.

Gardiner, B., Schuck, A., Schelhaas, M.-J., Orazio, C., Blennow, K., and Nicoll, B.: Living with Storm Damage to Forests What Science Can Tell Us What Science Can Tell Us, European Forest Institute, EFI, Joensuu, Finland, https://doi.org/10.13140/2.1.1730.2400, 2013.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, in: Deep Learning, pp. 164–223, MIT Press, 2016.

Govorushko, S. M.: Natural processes and human impacts: Interactions between humanity and the environment, Springer Science & Business Media, 2011.

Gregow, H., Laaksonen, A., and Alper, M. E.: Increasing large scale windstorm damage in Western, Central and Northern European forests, 1951-2010, Scientific Reports, 7, https://doi.org/10.1038/srep46397, 2017.

Guikema, S. D., Quiring, S. M., and Han, S. R.: Prestorm Estimation of Hurricane Damage to Electric Power Distribution Systems, Risk Analysis, 30, 1744–1752, https://doi.org/10.1111/j.1539-6924.2010.01510.x, 2010.

Guikema, S. D., Nateghi, R., Quiring, S. M., Staid, A., Reilly, A. C., and Gao, M.: Predicting Hurricane Power Outages to Support Storm Response Planning, IEEE Access, 2, 1364–1373, https://doi.org/10.1109/ACCESS.2014.2365716, 2014.

Haarsma, R. J., Hazeleger, W., Severijns, C., De Vries, H., Sterl, A., Bintanja, R., Van Oldenborgh, G. J., and Van Den Brink, H. W.: More hurricanes to hit western Europe due to global warming, Geophysical Research Letters, 40, 1783–1788, https://doi.org/10.1002/grl.50360, 2013.

Han, S. R., Guikema, S. D., and Quiring, S. M.: Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models, Risk Analysis, 29, 1443–1453, https://doi.org/10.1111/j.1539-6924.2009.01280.x, 2009.

Hanewinkel, M.: Neural networks for assessing the risk of windthrow on the forest division level: a case study in southwest Germany, European Journal of Forest Research, 124, 243–249, 2005.

Hanninen, K. and Naukkarinen, J.: 570 000 customers experienced power losses at the end of the year, (Loppuvuoden sähkökatkoista kärsi 570 000 asiakasta - Energiateollisuus, http://www.mynewsdesk.com/fi/pressreleases/loppuvuoden-saehkoekatkoista-kaersi-570-000-asiakasta-725038, 2012.

Hart, E., Sim, K., Kamimura, K., Meredieu, C., Guyon, D., and Gardiner, B.: Use of machine learning techniques to model wind damage to forests, Agricultural and forest meteorology, 265, 16–29, 2019.

He, J., Wanik, D. W., Hartman, B. M., Anagnostou, E. N., Astitha, M., and Frediani, M. E.: Nonparametric Tree-Based Predictive Modeling of Storm Outages on an Electric Distribution Network, Risk Analysis, 37, 441–458, https://doi.org/10.1111/risa.12652, 2017.

Hersbach, H., Bell, W., Berrisford, P., Horányi, A., J., M.-S., Nicolas, J., Radu, R., Schepers, D., Simmons, A., Soci, C., and Dee, D.: Global reanalysis: goodbye ERA-Interim, hello ERA5, ECMWF newsletter, pp. 17–24, https://doi.org/10.21957/vf291hehd7, https://www.ecmwf.int/node/19027, 2019.

Holdaway, M. R.: Spatial modeling and interpolation of monthly temperature using kriging, Climate Research, 6, 215–225, 1996.

Ilta-Sanomat: Rauli-myrsky repii Suomea: lähes 200000 taloutta ilman sähköjä - Kotimaa - Ilta-Sanomat, https://www.is.fi/kotimaa/art-2000001248969.html, 2016.

IPCC: Impacts of 1.5 C global warming on natural and human systems, Global warming of 1.5° C. An IPCC Special Report, 2018.

Jokinen, P., Vajda, A., and Gregow, H.: The benefits of emergency rescue and reanalysis data in decadal storm damage assessment studies, Advances in Science and Research, 12, 97–101, https://doi.org/10.5194/asr-12-97-2015, 2015.

Kankanala, P., Pahwa, A., and Das, S.: Regression models for outages due to wind and lightning on overhead distribution feeders, in: Power and Energy Society General Meeting, 2011 IEEE, pp. 1–4, IEEE, 2011.

Kankanala, P., Pahwa, A., and Das, S.: Estimation of Overhead Distribution System Outages Caused by Wind and Lightning Using an Artificial Neural Network, in: International Conference on Power System Operation & Planning, 2012.

Kankanala, P., Das, S., and Pahwa, A.: AdaBoost$^+$: An Ensemble Learning Approach for Estimating Weather-Related Outages in Distribution Systems, IEEE Transactions on Power Systems, 29, 359–367, 2014.

Karthick, S., Malathi, D., Sudarsan, J., and Nithiyanantham, S.: Performance, evaluation and prediction of weather and cyclone categorization using various algorithms, Modeling Earth Systems and Environment, pp. 1–9, 2020.

Kossin, J. P. and Sitkowski, M.: An objective model for identifying secondary eyewall formation in hurricanes, Monthly Weather Review, 137, 876–892, 2009.

Kron W., Schuck A.: After the floods, Tech. rep., Münchener Rückversicherungs-Gesellschaft (Munich Reinsurance Company), https://www.munichre.com/content/dam/munichre/global/content-pieces/documents/302-08121_en.pdf/_jcr_content/renditions/original./302-08121_en.pdf, 2013.

Kufeoglu, S. and Lehtonen, M.: Cyclone Dagmar of 2011 and its impacts in Finland, in: IEEE PES Innovative Smart Grid Technologies Conference Europe, vol. 2015-January, IEEE Computer Society, https://doi.org/10.1109/ISGTEurope.2014.7028868, 2015.

Lagerquist, R., McGovern, A., and Smith, T.: Machine learning for real-time prediction of damaging straight-line convective wind, Weather and Forecasting, 32, 2175–2193, 2017.

Lamb, H. and Knud, F.: Historic Storms of the North Sea, British Isles and Northwest Europe - Google Books, https://books.google.fi/books?hl=en&lr=&id=P4n1z9rOh5MC&oi=fnd&pg=PR8&ots=ewtYPSlQ1H&sig=VJJ00ehiPL0d1oaTDfa14nFb3zI&redir_esc=y#v=onepage&q=duration&f=false, 1991.

Li, Z., Singhee, A., Wang, H., Raman, A., Siegel, S., Heng, F.-L., Mueller, R., and Labut, G.: Spatio-temporal forecasting of weather-driven damage in a distribution system, in: 2015 IEEE Power & Energy Society General Meeting, pp. 1–5, IEEE, 2015.

Liu, Y., Zhong, Y., and Qin, Q.: Scene Classification Based on Multiscale Convolutional Neural Network, IEEE Transactions on Geoscience and Remote Sensing, 56, 7109 – 7121, https://doi.org/10.1109/TGRS.2018.2848473, https://ieeexplore.ieee.org/abstract/document/8421052, 2018.

Mäkisara, K., Katila, M., Peräsaari, J., and Tomppo, E.: The multi-source national forest inventory of Finland–methods and results 2013, 2016.

Masson-Delmotte, T., Zhai, P., Pörtner, H., Roberts, D., Skea, J., Shukla, P., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., et al.: IPCC, 2018: Summary for Policymakers. In: Global warming of 1.5 C. An IPCC Special Report on the impacts of global warming of 1.5 C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global, Geneva, Switzerland, 2018.

myrskyvaroitus.com: 2018 Syystrombeja, https://myrskyvaroitus.com/index.php/myrskytieto/myrskyhistoria/202-2018-syystrombeja, 2018.

Nateghi, R., Guikema, S., and Quiring, S. M.: Power Outage Estimation for Tropical Cyclones: Improved Accuracy with Simpler Models, Risk Analysis, 34, 1069–1078, https://doi.org/10.1111/risa.12131, 2014.

Niemelä, E.: KESKEYTYSTILASTO 2017 (i), Tech. Rep. 2018-06-14 11:51:52.916, Energiateollisuus Ry, Eteläranta 10, 00130 Helsinki, Finland, https://energia.fi/files/2785/Sahkon_keskeytystilasto_2017.pdf, 2018.

Nurmi, V., Pilli-Sihvola, K., Gregow, H., and Perrels, A.: Overadaptation to Climate Change? The Case of the 2013 Finnish Electricity Market Act, Economics of Disasters and Climate Change, 3, 161–190, https://doi.org/10.1007/s41885-018-0038-1, 2019.

Pellikka, P. and Järvenpää, E.: Forest stand characteristics and wind and snow induced forest damage in boreal forest, in: Proceedings of the International Conference on Wind Effects on Trees, held in September, pp. 16–18, Citeseer, 2003.

Peltola, H., Kellomäki, S., and Väisänen, H.: Model Computations of the Impact of Climatic Change on the Windthrow Risk of Trees, Climatic Change, 41, 17–36, https://doi.org/10.1023/A:1005399822319, 1999.

Pinto, J. G., Bellenbaum, N., Karremann, M. K., and Della-Marta, P. M.: Serial clustering of extratropical cyclones over the North Atlantic and Europe under recent and future climate conditions, Journal of Geophysical Research: Atmospheres, 118, 12,476–12,485, https://doi.org/10.1002/2013JD020564, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JD020564, 2013.

Platt, J. et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers, 10, 61–74, 1999.

Ramon, J., Lledó, L., Torralba, V., Soret, A., and Doblas-Reyes, F. J.: What global reanalysis best represents near-surface winds?, Natural Disasters and Adaptation to Climate Change, 145, 3236–3251, https://doi.org/10.1002/qj.3616, 2019.

Rasmussen, C. E.: Gaussian processes in machine learning, in: Summer School on Machine Learning, pp. 63–71, Springer, 2003.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.02881, 2017.

Rossi, P. J.: Object-Oriented Analysis and Nowcasting of Convective Storms in Finland, Ph.D. thesis, Aalto University, http://urn.fi/URN:ISBN:978-952-60-6441-3, 2015.

Rupp, M.: Machine learning for quantum mechanics in a nutshell, International Journal of Quantum Chemistry, 115, 1058–1073, 2015.

Schelhaas, M.-J.: Impacts of natural disturbances on the development of European forest resources: application of model approaches from tree and stand levels to large-scale scenarios, Dissertationes Forestales, https://doi.org/10.14214/df.56, 2008.

Schelhaas, M.-J., Nabuurs, G.-J., and Schuck, A.: Natural disturbances in the European forests in the 19th and 20th centuries, Global Change Biology, 9, 1620–1633, https://doi.org/10.1046/j.1365-2486.2003.00684.x, https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2486.2003.00684.x, 2003.

Seidl, R., Schelhaas, M., Rammer, W., and Verkerk, P.: Increasing forest disturbances in Europe and their impact on carbon storage, Nature Climate Change, 4, 806–810, https://doi.org/10.1038/NCLIMATE2318, 2014.

Shawe-Taylor, J., Cristianini, N., et al.: Kernel methods for pattern analysis, pp. 296–297, Cambridge university press, 2004.

Shield, S. et al.: Predictive Modeling of Thunderstorm-Related Power Outages, Master's thesis, The Ohio State University, 2018.

Singhee, A. and Wang, H.: Probabilistic forecasts of service outage counts from severe weather in a distribution grid, in: 2017 IEEE Power & Energy Society General Meeting, pp. 1–5, IEEE, 2017.

Suvanto, S., Henttonen, H. M., Nöjd, P., and Mäkinen, H.: Forest susceptibility to storm damage is affected by similar factors regardless of storm type: Comparison of thunder storms and autumn extra-tropical cyclones in Finland, Forest Ecology and Management, 381, 17 – 28, https://doi.org/https://doi.org/10.1016/j.foreco.2016.09.005, http://www.sciencedirect.com/science/article/pii/S0378112716305266, 2016.

Tervo, R., Karjalainen, J., and Jung, A.: Short-Term Prediction of Electricity Outages Caused by Convective Storms, IEEE Transactions on Geoscience and Remote Sensing, 57, 8618–8626, 2019.

Ukkonen, P. and Mäkelä, A.: Evaluation of machine learning classifiers for predicting deep convection, Journal of Advances in Modeling Earth Systems, 11, 1784–1802, 2019.

Ukkonen, P., Manzato, A., and Mäkelä, A.: Evaluation of thunderstorm predictors for Finland using reanalyses and neural networks, Journal of Applied Meteorology and Climatology, 56, 2335–2352, 2017.

Ulbrich, U., Pinto, J. G., Kupfer, H., Leckebusch, G., Spangehl, T., and Reyers, M.: Changing Northern Hemisphere storm tracks in an ensemble of IPCC climate change simulations, Journal of Climaste, 21, 1669–1679, 2008.

Ulbrich, U., Leckebusch, G. C., and Pinto, J. G.: Extra-tropical cyclones in the present and future climate: A review, in: Theoretical and Applied Climatology, vol. 96, pp. 117–131, Springer Wien, https://doi.org/10.1007/s00704-008-0083-8, 2009.

Valta, H., Lehtonen, I., Laurila, T., Venäläinen, A., Laapas, M., and Gregow, H.: Communicating the amount of windstorm induced forest damage by the maximum wind gust speed in Finland, Advances in Science and Research, 16, 31–37, https://doi.org/10.5194/asr-16-31-2019, 2019.

Virot, E., Ponomarenko, A., Dehandschoewercker, Quéré, D., and Clanet, C.: Critical wind speed at which trees break, Physical Review E, 93, https://doi.org/10.1103/PhysRevE.93.023001, 2016.

Wang, G., Xu, T., Tang, T., Yuan, T., and Wang, H.: A Bayesian network model for prediction of weather-related failures in railway turnout systems, Expert Systems with Applications, 69, 247–256, https://doi.org/10.1016/j.eswa.2016.10.011, 2017.

Weather and Safety Center of Finnish Meteorological Institute - Duty forecasters: Personal communication, 05/2020.

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, pp. 76–99, Academic press, 2011.

Williams, C. K. and Rasmussen, C. E.: Gaussian processes for machine learning, vol. 2, MIT press Cambridge, MA, 2006.

Yang, F., Wanik, D. W., Cerrai, D., Bhuiyan, M. A. E., and Anagnostou, E. N.: Quantifying uncertainty in machine learning-based power outage prediction model training: A tool for sustainable storm restoration, Sustainability, 12, 1525, 2020.

Yue, M., Toto, T., Jensen, M. P., Giangrande, S. E., and Lofaro, R.: A Bayesian approach-based outage prediction in electric utility systems using radar measurement data, IEEE Transactions on Smart Grid, 9, 6149–6159, https://doi.org/10.1109/TSG.2017.2704288, 2018.

Predicting power outages caused by extratropical storms

Roope Tervo[1,*], Ilona Láng[1,*], Alexander Jung[2], and Antti Mäkelä[1]

[1]Finnish Meteorological Institute, B.O. 503, 00101 Helsinki, Finland
[2]Aalto University, Dept of Computer Science, B.O. 11000, 00076 Aalto, Finland
[*]These authors contributed equally to this work.

Correspondence: Roope Tervo (roope.tervo@fmi.fi)

Abstract. Strong winds induced by extratropical storms cause a large number of power outages, especially in highly forested countries such as Finland. Thus, predicting the impact of the storms is one of the key challenges for power grid operators. This article introduces a novel method to predict the storm severity for the power grid employing ERA5 reanalysis data combined with forest inventory. We start by identifying storm objects from wind gust and pressure fields by using contour lines of 15 $\mathrm{m\,s^{-1}}$ and 1000 hPa respectively. The storm objects are then tracked and characterized with features derived from surface weather parameters and forest vegetation information. Finally, objects are classified with a supervised machine learning method based on how much damage to the power grid they are expected to cause. Random Forest Classifier, Support Vector Classifier, Naive Bayes, Gaussian Processes, and Multilayer Perceptron were evaluated for the classification task, Support Vector Classifier providing the best results.

1 Introduction

Strong winds ~~,~~ caused by extratropical storms are among the ~~biggest~~ most significant natural hazards in Europe, causing massive damage to the forests and society (~~i.e.~~e.g. Schelhaas et al. (2003); Schelhaas (2008); Ulbrich et al. (2008); Seidl et al. (2014); Valta et al. (2019)); extratropical storms are responsible for 53 percent of all losses related to natural hazards in Europe (Kron W., Schuck A., 2013). Such storms pose a huge challenge for power distribution companies in highly-forested countries such as Finland (Gardiner et al., 2010) where falling trees cause power outages for hundreds of thousands of customers every year (Niemelä, 2018). ~~Having~~ The windstorms create a significant risk for the power supply in Finland, which has over 90 000 kilometers ~~overhead line~~ of overhead lines (70 percent of it medium-voltage, 1-35 kV, network) passing through forest (Kufeoglu and Lehtonen, 2015)~~, the windstorms create significant risk for the power supply in Finland~~. Between the years 2010 and 2018, on average 46 percent of all transmission faults in Finland were caused by ~~extra-tropical~~ extratropical storms (Finnish Energy, 2010-2018). During the years of the most damaging storms, 2011 and 2013, the share of windstorm damages of all fault causes was up to 69 percent (Finnish Energy, 2011, 2013)~~compared to previous years~~. The need for managing power interruptions is even more urgent since the power suppliers in Finland are obliged to financially compensate customers

of urban areas after 6 hours and rural areas after 36 hours of interruption in electricity distribution (Nurmi et al., 2019)~~, thus~~. Thus they require a large amount of ~~manpower~~ workforce to fix caused damages rapidly.

~~Based on Gregow et al. (2017), the windstorm damages especially in Northern, Central, and Western Europe have increased during the past three decadessignificantly~~As Ulbrich et al. (2009) describe, there is no scientific consensus on how the occurrence and magnitude of extratropical storms will evolve in the future. Based on existing literature, the windstorm-related damages are increasing, while it remains unclear whether this is due to the higher exposure of society or the number and intensity of extratropical storms. Gregow et al. (2017) discovered that windstorm damages had increased significantly during the previous three decades, especially in northern, central, and western Europe. Also, ~~other studies are suggesting~~ several other studies suggest an increase in wind-related damages in Europe (Csilléry et al. (2017); Haarsma et al. (2013); Gardiner et al. (2010)). ~~Although Ulbrich et al. (Ulbrich et al., 2009) describe the future of extratropical storms to be complex to foresee, it seems that~~ Interestingly, some studies detected a decrease in the total number of ~~storms might decrease~~ extratropical storms (i.e. Donat et al. (2011))~~but on~~ , while others found an increase in the number of extreme storms in specific regions, like western Europe and ~~Northeast Atlantic , the number of extreme storms increases (e. g. Pinto et al. (2013)). Besides, the tracks of extratropical storms have already been shifted and are likely to shift also in the future~~ northeast Atlantic (Pinto et al., 2013). Another supporting view of a potential increase in extratropical storms in northern Europe can be found in the IPCC (2018) report. The report states that extratropical storm tracks are shifting towards the poles~~(IPCC, 2018)~~, which might affect the storminess in northern Europe. ~~According to (Barredo, 2010) the increased disaster losses to be caused rather by~~ Thus, it may be concluded that also the losses related to extratropical storms are likely to increase especially in northern Europe. However, as Barredo (2010) emphasizes, the cause for increased losses can at least partly be explained by the increasing exposure of society rather than the increased number of windstorms.

Several previous studies respond to the demand for storm impact estimation for power distribution, many of them focusing on the hurricane-induced power blackouts in ~~Northern~~ northern America (Eskandarpour and Khodaei (2017); Guikema et al. (2014, 2010); Nateghi et al. (2014); Han et al. (2009); Wang et al. (2017); Allen et al. (2014); Chen and Kezunovic (2016); He et al. (2017); Liu et al. (2018)). Convective thunderstorms have ~~been also~~ also been investigated thoroughly. Li et al. (2015) introduced an area-based outage prediction method further developed to take power grid topology into account (Singhee and Wang, 2017). Shield et al. (2018) studied outage prediction by applying a random forest classifier to weather forecast data in a regular grid. Kankanala et al. used data from ground observation stations and experimented regression (Kankanala et al., 2011), a multilayer perceptron neural network (Kankanala et al., 2012), and ensemble learning (Kankanala et al., 2014) to predict outages caused by wind and thunder. The Bayesian outage probability (BOP) prediction model ~~(Yue et al., 2018)~~ developed by Yue et al. (2018) combines weather radar data and unifies it to a regular grid. Cintineo et al. (2014) create spatial objects from satellite and weather radar data, and track and classify the objects with the Naïve Bayesian classifier. Rossi (2015) developed a method to detect and track convective storms. The method was ~~later~~ further developed to predict power outages (Tervo et al., 2019).

While much work exists on damage caused by ~~large-scale storms (hurricanes) and small-scale storms (convective thunderstorms), relatively little has been done to be prepared for~~ hurricanes and convective thunderstorms, relatively few examples exist relating

to outages caused by mid-latitude extratropical storms differing from hurricanes and convective storms in available data, time-span, and applicable methods for detecting and tracking. Extratropical storms are considered, for example, in Yang et al. (2020), where different decision tree methods are applied to a regular grid in the outage prediction task. Cerrai et al. (2019) also uses decision trees and regular grid for the outage prediction taking tree-leaf conditions into account as a predictive feature. Related forest damage studies have been conducted ~~, though,~~ with random forest classifiers and neural networks. Hart et al. (2019) showed ~~at~~ that random forest regression and artificial neural networks ~~can predict a~~ could predict the number of falling trees in France caused by the wind. Hanewinkel (2005) conducted a similar study in Germany using artificial neural networks. Artificial neural networks have been used to predict extreme weather in Finland ~~(Ukkonen et al. (2017), Ukkonen and Mäkelä (2019)). To summarise, according to various sources, for example, the~~ (Ukkonen et al., 2017; Ukkonen and Mäkelä, 2019). The framework of IPCC ~~(Masson-Delmotte et al., 2018),~~ (2018) emphasizes that the impacts of ~~the~~ extreme weather risks can be analyzed by estimating the hazard, vulnerability, and exposure~~while machine learning techniques are becoming more popular in the task of connecting the natural hazards~~. In an increasing manner, connecting these fields (i.e., the natural hazard with the societal ~~impact forecasts~~ factors) is done with machine learning (Chen et al., 2008).

We present a novel method to identify, track, and classify extratropical storm objects based on how much power outages they are expected to induce. We adapt convective storm object detection (Rossi (2015), Tervo et al. (2019), Cintineo et al. (2014)) to find potentially harmful areas from extratropical storms by contouring objects from pressure and wind gust fields. ~~We then~~ Instead of highly-localized convective storms, we aim at larger but still regional geospatial accuracy so that, for example, damages in western and eastern Finland can be distinguished. We train a supervised machine learning model to classify storm objects according to their damage potential. To our knowledge, our method is the first that employs the extratropical storm objects as polygons and combines them with meteorological and non-meteorological features to predict power outages. The method can be used as a decision support tool in power distribution companies or as part of elaborating impact forecast by duty forecasters in national hydro-meteorological centers. The ERA5 atmospheric reanalysis (European Centre for Medium-Range Weather Forecasts, 2017) provides the primary meteorological input data for this study, while the national forest inventory provided by The Natural Resources Institute Finland (Luke) is used to represent the forest conditions in the prediction. Finally, historically occurred power outages from two sources are used to train the model. However, the operational use of the model would require the use of weather prediction data instead of reanalysis.

This paper is organized as follows: Chapter 2 presents ~~used data and followed by~~ the used data, which is followed by a step-by-step method description in Chapter 3. Chapter 3.1 discusses identifying storm objects and ~~present~~ explains the storm tracking algorithm. Chapter 3.2 ~~discusses~~ considers storm and forest characteristics, hereafter called features. Chapter 3.3 discusses how to define labels of storm objects based on the outage data. Chapter 3.4 describes the used machine learning methods. In Chapter 4, we discuss the performance of the method~~followed by conclusion in Chapter 5~~. Finally, Chapter 5 includes a discussion and conclusions.

2 Data

We base our method on three main data sources: ERA5 reanalysis data (Hersbach et al., 2019), multi-source national forest inventory (ms-nfi) provided by The Natural Resources Institute Finland (Luke), and occurred power outages obtained from two sources~~.~~: First, the *local dataset* is gathered from two power distribution companies, Loiste and Järvi-Suomen Energia (JSE), located in Eastern Finland. Second, the *national dataset* is obtained from Finnish Energy (ET), a branch organization for the industrial and labor market policy of the energy sector. All data ~~is gathered~~ consider years from 2010 to 2018. ~~These data are described in the following.~~

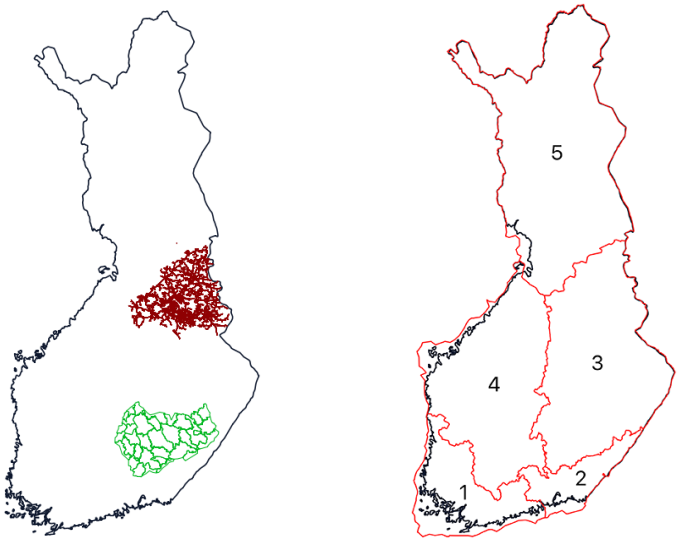
ERA5 is the newest generation reanalysis data provided by ECMWF. ERA5 covers the years from 1979 onward with a one-hour temporal resolution, has a horizontal resolution of 31 km, and covers the atmosphere using 137 levels up to a height of 80 km (Hersbach et al., 2019). Compared to in-situ wind observations, reanalysis data provides a spatiotemporally wider dataset. However, a question may arise about the accuracy of the reanalysis data. Ramon et al. (2019) examined the wind speed characteristics of a total of five state-of-the-art global reanalyses concerning 77 instrumented towers. In their study, ERA5 had the best agreement with in-situ observations on daily time scales; this suggests the ERA5 wind parameters to be ~~decent~~ adequate in windstorm damage examinations as well. ERA5 data are also known to contain unrealistically large surface wind speeds in some locations (European Centre for Medium-Range Weather Forecasts, 2019). None of these locations are, nevertheless, inside the geographical domain of this work.

The multi-source forest inventory data is based on field measurements, satellite observations, digital maps, and other geo-referenced data sources (Mäkisara et al., 2016). The data consists of estimates for the forest age, tree species dominance, the mean and total volume, and the biomass (total and tree species-specific). The original geospatial resolution of the data is 16 meters, which has been reduced to approximately 1.6 km resolution to speed up the processing. Taking into account the size of extratropical cyclones (diameter 1000 km) and the wide areas where wind damages typically occur e.g. near to the cold front, we consider a resolution of 1.6 km being sufficiently high for modeling wind storm damages.

~~Each power distribution company in Finland is monitoring and collecting the power interruption data. The raw data consist of many parameters, including the start and end times of~~ Power outage data are obtained from two complementary sources. *The national dataset* is acquired from the ~~interruption along with a location of a distribution transformer. This data is collected by the~~ Finnish Energy ~~(Finnish Energy, 2010-2018)~~(2010-2018) who aggregates the data from power distribution companies in Finland. The national data ~~can be acquired~~ are provided only for research purposes ~~but only~~ and for areas containing a minimum of ~~five~~ six grid companies; this is, for example, to ensure ~~the anonymity of energy users. The same data can be obtained also from some~~ energy users' anonymity. Therefore, the national dataset does not include exact locations of the faults. We have also obtained some parts of the data with better spatial accuracy from two individual power distribution companies~~with better spatial accuracy.~~. In this paper, we refer to this data as *the local dataset*. In the local dataset, the fault locations are reported in relation to transformers, i.e. the spatial resolution of the outages ranges from a few meters to kilometers.

Figure 1 illustrates the geographical coverage of the power outage data. The local dataset contains all outages from 2010 to 2018 in the northern area (Loiste) and outages related to major storms in the southern area (JSE), shown in Figure 1a.

The national dataset contains all outages in Finland from 2010 to 2018 divided into five regions, shown in Figure 1b. The national dataset contains in total 6 140 434 outages with relatively low geographical accuracy. On the other hand, the local dataset represents a substantially smaller geographical area with a good geographical accuracy but contains only 22 028 outages in total. We train our classification models, described in more detail in Chapter 3.4, with both datasets to evaluate their performance for different types of data.



(a) (b)

Figure 1. (a) Geographical coverage of the outage data (local dataset). The red lines represent the power grid of Loiste (northern grid company) and the green lines the operative areas of JSE (southern grid company). Outages of the local dataset are collected from both areas. (b) Regions in the national outage dataset. Outages are gathered from entire Finland and aggregated to the regions shown in the figure.

3 Method

We predict power outages by classifying storm objects identified from gridded weather data into three classes based on ~~a~~ the number of power outages the storm ~~can typically cause~~typically causes. The overall process ~~contains~~ consists of the following steps: (1) identifying storm objects from weather fields by finding contour lines of ~~some particular threshold~~particular thresholds, (2) tracking the storm object movement, (3) gathering features of the storm objects, and (4) classifying ~~the objects.~~ each storm object individually. The classification is conducted for each storm object separately to distinguish the different damage potential. Tracking is, however, necessary to gather necessary features such as object movement speed and direction. In the following, we discuss these phases in more detail.

3.1 Identifying and tracking storm objects

Storm objects are identified by finding contour lines of 10-meter wind gust fields ~~and pressure fields with several thresholds. We define the storm objects~~ using 15 $\mathrm{m\,s^{-1}}$ thresholds from the ERA5 surface level grid with a time step of 1 hour. The contouring algorithm is capable of finding interior rings of the polygons. The used wind gust fields did not, however, contain such cases. Thus one storm object represents a solid area (polygon) where the hourly maximum wind gust exceeds 15 $\mathrm{m\,s^{-1}}$ during one particular hour. The threshold of 15 $\mathrm{m\,s^{-1}}$ is selected as different sources indicate Finland being vulnerable for windstorms and rather moderate winds (from 15 $\mathrm{m\,s^{-1}}$) causing damages to forests (Valta et al., 2019; Gardiner et al., 2013). ~~To~~ Valta et al. (2019) developed a method to estimate the windstorm impacts on forests ~~, Valta et al. (2019) developed a method~~ by combining the recorded forest damages from the nine most intense storms and their observed maximum inland wind gusts. According to the formula developed in the study, ~~alone~~ the inland wind gusts of 15 $\mathrm{m\,s^{-1}}$ alone result in forest damages of 1800 $\mathrm{m^3}$. We also identify pressure objects by finding contour lines using a 1000 hPa threshold to connect potentially distant storm objects around the low-pressure center to the same storm event.

After identification, storm objects are ~~connected to preceding objects using Algorithm 1. Each object having pressure objects or preceding~~ tracked by connecting them with each other. Each storm object is first connected to nearby pressure objects from the current and preceding time steps. If pressure objects do not exist within the distance threshold, the object is connected to nearby storm objects from the current and preceding time steps. The Algorithm enables assigning each storm object to an overall event (low pressure system) and tracking the objects~~within the threshold, are assigned to the same storm event and gets the same storm ID. Notably, nearby pressure objectsare considered along with previous wind objects in assigning ID. This allows several, potentially distant wind objects around the low-pressure center to be assigned to the same storm event. Single wind objects without nearby pressure object or preceding objects are left without ID as they are not assumed to be part of any storm.~~' movement. Algorithm 1 shows the details of the process.

We use a 500 km distance threshold for the distance between ~~wind~~ the storm and pressure objects. As the typical diameter of an extratropical storm is approximately 1000 km (Govorushko, 2011), we assume ~~, the damaging wind~~ the damaging storm objects to situate a maximum 500 km from the center of the low pressure. The ~~thresholds for motion speed for wind objects~~ threshold for movement speed is 200 $\mathrm{km\,h^{-1}}$ for storm objects and 45 $\mathrm{km\,h^{-1}}$ for pressure objects. ~~Wind objects are i.e.~~ In other words, storm objects are not assumed to move ~~over~~ more than 200 km and pressure objects ~~over~~ more than 45 km from the preceding hourly time step (Govorushko, 2011). Convective storms may move faster but are outside the focus of this work.

3.2 Extracting storm object features

We characterize the storm objects identified by the methods discussed in Section 3.1 using the features listed in Table 1. The features are structured as four groups. The first group is a number of object characteristics such as size and movement speed and direction, which are calculated from the contoured storm objects themselves. As ~~a~~ the second group, relevant weather conditions, such as wind speed, temperature, ~~etc.~~and others, are extracted from ERA5 data. ~~To~~ We aggregate values as a minimum, maximum, average, and standard deviation calculated over all grid cells under the object coverage to represent each

Algorithm 1 Storm tracking

Input

~~Individual storm objects~~ Storm and pressure objects S_o arranged by time

$pressure\ distance\ threshold$

$wind\ distance\ threshold$

$speed\ threshold$

$time\ step$

Output

Connected storm ~~object~~ and pressure objects with storm ID

for all storm and pressure object $O_{w|p} \in S_o$ **do**

$current\ time \leftarrow$ time of the object $O_{w|p}$

$previous\ time \leftarrow current\ time - time\ step$

Current time pressure objects $S_p^c \leftarrow$ pressure objects having centroid within $pressure\ distance\ threshold$ from object $O_{w|p}$ centroid and time stamp $current\ time$

Previous time pressure objects $S_p^p \leftarrow$ pressure objects having centroid within $speed\ threshold$ from .8mm object $O_{w|p}$ centroid and time stamp $previous\ time$

Current time storm objects $S_w^c \leftarrow$ storm objects having centroid within $wind\ distance\ threshold$ from .8mm object $O_{w|p}$ centroid and time stamp $current\ time$

Previous time storm objects $S_w^p \leftarrow$ storm objects having centroid within $speed\ threshold$ from .8mm object $O_{w|p}$ centroid and time stamp $previous\ time$

if pressure object $O_p^c \in S_p^c$ exists with ID **then**

Use pressure object ~~ID~~ O_p^c ID

else if pressure object $O_p^p \in S_p^p$ exists with ID **then**

Use previous ~~pressure object ID~~ time pressure object O_p^p ID

else if storm object $O_w^c \in S_w^c$ exists with ID **then**

Use ~~other object ID~~ storm object O_w^c ID

else if storm object $O_w^p \in S_w^p$ exists with ID **then**

Use previous ~~object ID~~ time storm object O_w^p ID

else if storm or pressure object $O_{w|p}^p \in S_w^p \cup S_p^p$ exists without ID **then**

Give new ID to the previous object $O_{w|p}^p$ and current object $O_{w|p}$

else

Leave object $O_{w|p}^p$ without ID

end if

end for

parameter with one number~~, we aggregate values from the object coverage using functions listed in Table 1~~. Third, as most of the outages are caused by the trees falling ~~over~~ on power grid lines (Campbell and Lowry, 2012), ~~and thus the features~~ the characteristics of the forest ~~(i.e. tree height, age, or specie) contribute in~~ contribute to the damages (Peltola et al., 1999), we ~~support~~ complement our data with forest information. As ~~with~~ for weather parameters, values are aggregated ~~from~~ over the storm object coverage. The fourth group consists of the number of outages and affected customers used as labels in the model training process discussed ~~more~~ in more detail in Chapter 3.4.

~~We first gather several parameters and further select the most relevant ones.~~ We selected the 35 parameters based on two main criteria: First, we prepared a list of potential parameters detected in related studies, e.g. Suvanto et al. (2016); Peltola et al. (1999); Valta et a , or identified through the empirical experience of duty forecasters (Weather and Safety Center of Finnish Meteorological Institute - Duty fo . Second, we selected the relevant parameters, which were available to us or accessible with a reasonable effort. However, some possibly essential parameters, like soil temperature from ERA5 reanalysis, were left out because of the slow downloading process.

After the preliminary selection of the parameters, we conducted dozens of light experiments using different combinations of parameters and models to find the best possible setup. To this end, we ~~plot the difference in fitted Gaussian distribution between all samplesand class one and two samples~~fitted the Gaussian distribution to each parameter using at first all samples, then samples with few outages, and finally with many outages (classes 1 and 2 specified in Section 3.3). While many other distributions are known to suit better in modeling particular parameters~~(~~, such as Gamma in precipitation, Weibull in wind speed, and Lognormal in cloud properties (Wilks, 2011)~~),~~ , the Gaussian distribution is a sufficient simplification to help in selecting relevant parameters. ~~Distribution of~~ We visually inspected the differences between fitted Gaussian distributions to deduce the potential relevance of the parameter. Supposedly the distribution of one parameter is different for all samples and samples with many outages, and the classification method may exploit the parameter to predict the damage potential of the storm object. The distributions of some selected parameters ~~is shown in Appendices A1 and A1~~are shown in Appendix A. In total, 35 parameters, shown as ~~bolded~~ boldfaced in Table 1, were chosen for the final classification.

3.3 Defining classes

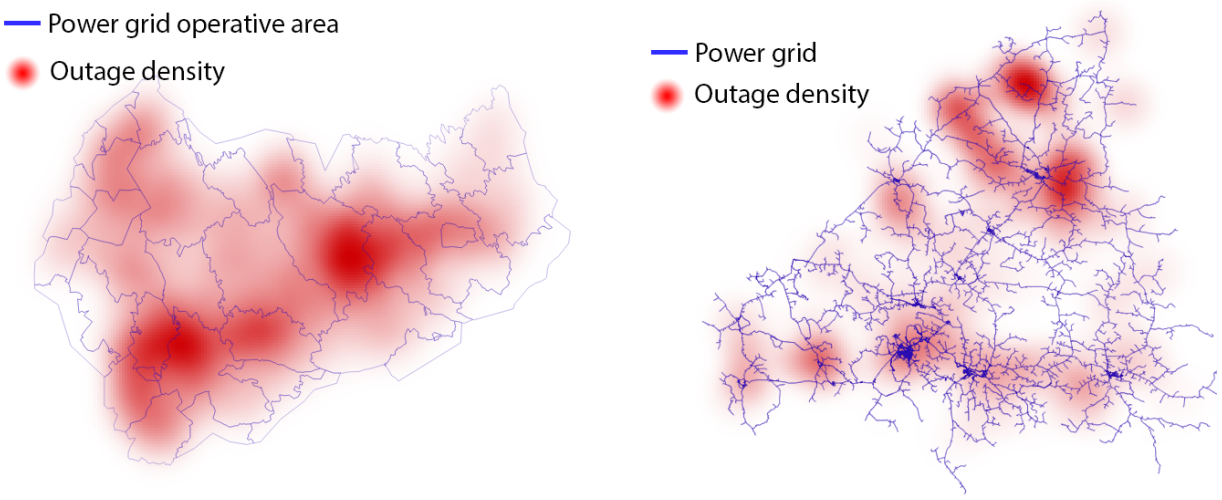
~~We use three classes designed together with power grid companies aiming at a simple "at glance" view for power grid operators. Class 0 represents no damage, class 1 low damage, and class 2 high damage. Next, we discuss the power outage data used in this project and then define limits for the classes.~~

~~The geographical coverage of the power outage data is illustrated in Figures 1a and 1b. The local dataset contains all outages from Northen Area (Loiste) and outages related to major storms in the Southern area (JSE). The national dataset contains all outages in Finland divided into five regions shown in Figure 1b.~~ As shown in Figures ~~?? and ??~~2a and 2b, the outages in the local dataset are concentrated heavily on 'hot-spots', assumingly, due to forest characteristics and network topology. ~~In total, the~~ The local dataset contains ~~24 542~~ 24 542 storm objects and ~~5 837 outages attributed to a storm~~ 5 837 outages connected to 2 363 storm objects. Thus 22 179 storm objects in the local dataset did not cause any outages. The local power outage data contain 16 191 outages, which can not be connected to any storm object. The national dataset contains ~~142 873~~ 142 873 storm

Table 1. Extracted features. Features used in the final classification marked as bold.

Feature	Aggregation	Explanation
Speed	-	Object movement speed
Angle	-	Object movement angle
Area	-	Object size
Area difference	-	Object area difference to the previous time step
Week	-	Week of the year
Snowdepth	average, minimum, maximum	Snow depth
Total column water vapor	**average, minimum, maximum**	Total amount of water vapour
Temperature	**average, minimum, maximum**	2 meter air temperature
Snowfall	average, minimum, maximum, sum	Snowfall (meter of water equivalent)
Total cloud cover	average, minimum, maximum	Total cloud cover (0-1)
CAPE	**average**, minimum, **maximum**	Convective available potential energy (J/kg)
Precipitation kg/m2	average, minimum, maximum, sum	Precipitation amount (kg/m2)
Wind gust	**average**, minimum, **maximum, standard deviation**	Hourly maximum wind gust ($\mathrm{m\ s^{-1}}$)
Wind Speed	**average**, minimum, **maximum, standard deviation**	10 meter wind speed ($\mathrm{m\ s^{-1}}$)
Wind Direction	**average**, minimum, maximum, **standard deviation**	Wind direction (degrees))
Dewpoint	**average, minimum, maximum**	Dewpoint)
Mixed layer height	**average, minimum, maximum**	Boundary layer height
Pressure	average, **minimum**, maximum	Air pressure
Forest age	**average**, minimum, maximum, standard deviation	The age of the growing stock on a forest stand
Forest site fertility	**average**, minimum, maximum, standard deviation	Group of the forest by vegetation zones
Forest stand mean diameter	**average**, minimum, maximum, standard deviation	Forest stand mean mean diameter
Forest stand mean height	**average**, minimum, maximum, standard deviation	Forest stand mean height
Forest canopy cover	**average**, minimum, maximum, standard deviation	Forest canopy cover fraction (0-100%)
Outages	-	Number of occured outages
Customers	-	Number of affected customers
Tansformers	-	Number of transformers under the object
All customers	-	Number of customers under the object
Class	-	Assigned class

objects and ~~5 965 324 outages attributed to some storm~~ 5 965 324 outages connected to 33 796 storm objects. 109 077 storm objects are not connected to any outages, and 175 110 outages can not be connected to any storm object.



(a) (b)

Figure 2. ~~(a) Geographical coverage~~ Spatial distribution of the outages ~~in local dataset~~between 2010 and 2018 visualised as a spatial heatmap. ~~Red lines represents the power grid where outages are gathered from. (b~~a) ~~Regions in national outage dataset. Outages are gathered from the whole Finland but aggregated to the regions shown in the image. (c) Spatial distribution of the outages in the~~ JSE ~~Network~~ network (~~Southern~~ southern area) ~~, data gathered between 2010 and 2018. (d~~b) ~~Spatial distribution of the outages in the~~ Loiste ~~Network~~ network (~~Northern~~ northern area)~~, data gathered between 2010 and 2018.~~

It ~~is notable,~~ should be noticed that the damage may occur anywhere in the power grid. Outages are, however, always reported as transformers without electricity. Typically one physical damage between the transformers causes several transformers to lose power. Power grid operators can often ~~afterward~~ turn part of the transformers back to operation even before fixing the ~~original~~ damage~~. This~~ actual damage, which causes an unavoidable noise to the datasets.

~~Figure 3 depicts~~ Figure 3 represents the number of outages and storm objects in both ~~,~~ local and national datasets. We can identify a large amount of 15 $\mathrm{m\ s^{-1}}$ storm objects in both sets, indicating that moderate wind ~~objects~~ without other influencing factors ~~, do not cause damage for~~ does not damage the transformers. When identifying storm objects with the contour of 20 and 25 $\mathrm{m\ s^{-1}}$, the number of objects reduces and ~~correlates~~ starts to correlate more with a high number of outages~~. This~~ , which supports views of previous studies showing the significance of stronger wind gusts to ~~higher~~ more severe storm damages. The method seems to identify also the most ~~important~~ critical storm days by capturing several storm objects for those days. For instance, at the end of 2013~~when~~ , when the three major storms Eino, Oskari, Seija (Valta et al., 2019) hit Finland, both datasets contain plenty of ~~wind objects with~~ storm objects with the 20 $\mathrm{m\ s^{-1}}$ threshold. Nevertheless, our experiments indicated that employing 15 $\mathrm{m\ s^{-1}}$ storm objects yielded the best results. This is described more in Chapter 4.

Figure 4 illustrates how much outages a single storm object typically produces. In the local dataset, most of the storm objects cause only a few outages. Only 65 storm objects, which are only 0.3 percent of the whole dataset, induced more than ~~10~~ ten outages. On the other hand, in the national dataset where one storm object typically affects several different transformers, ~~17~~

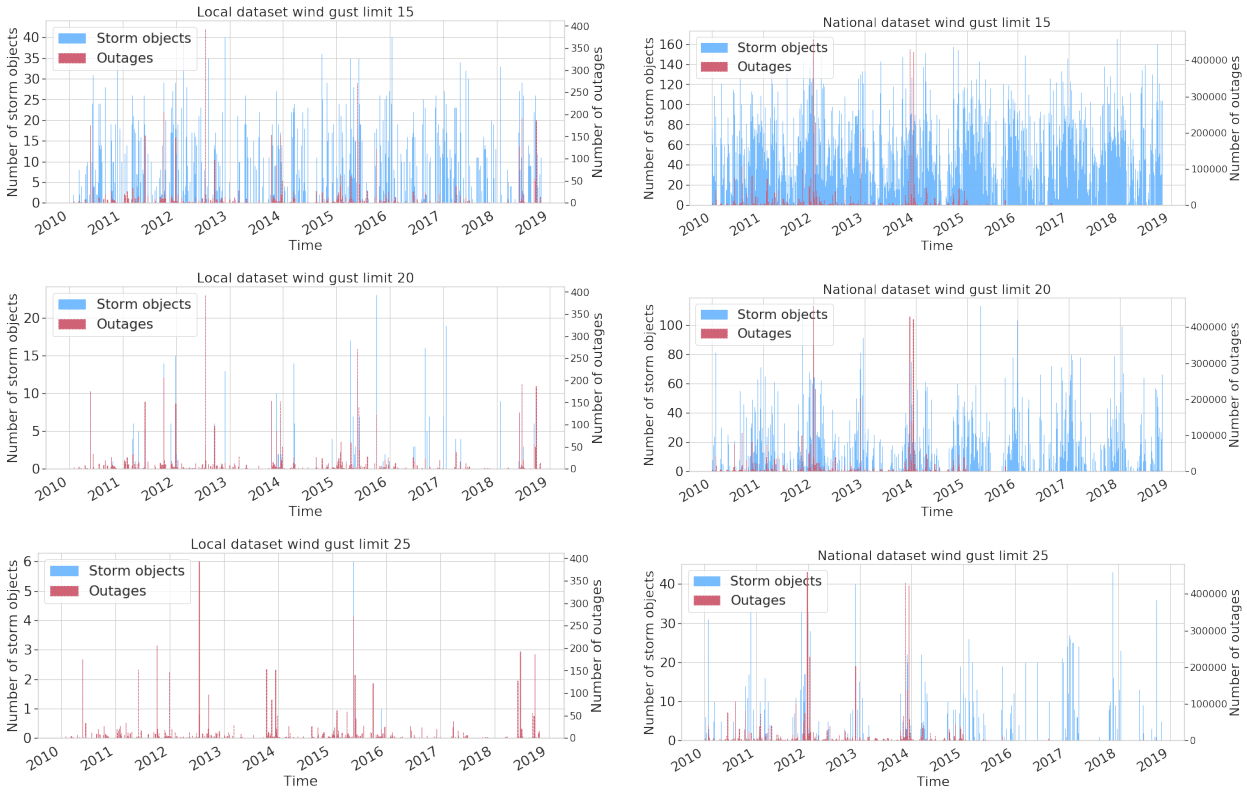
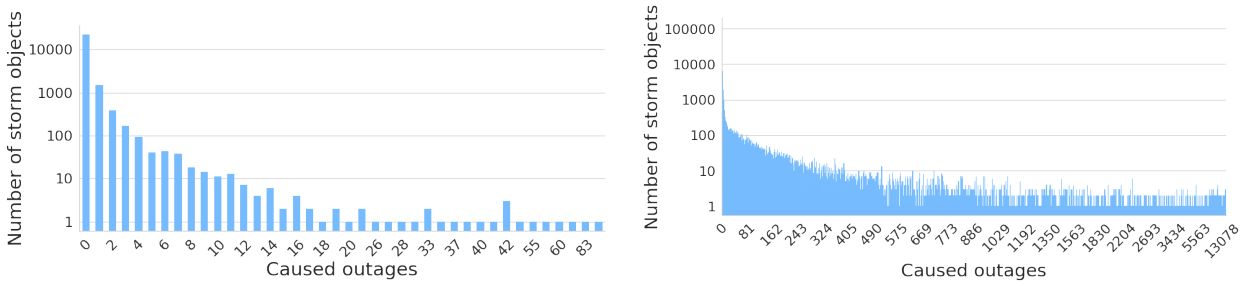


Figure 3. Storm object time series (15, 20 and 25 $\mathrm{m\ s^{-1}}$ contours) with occurred outages for local and national datasets.

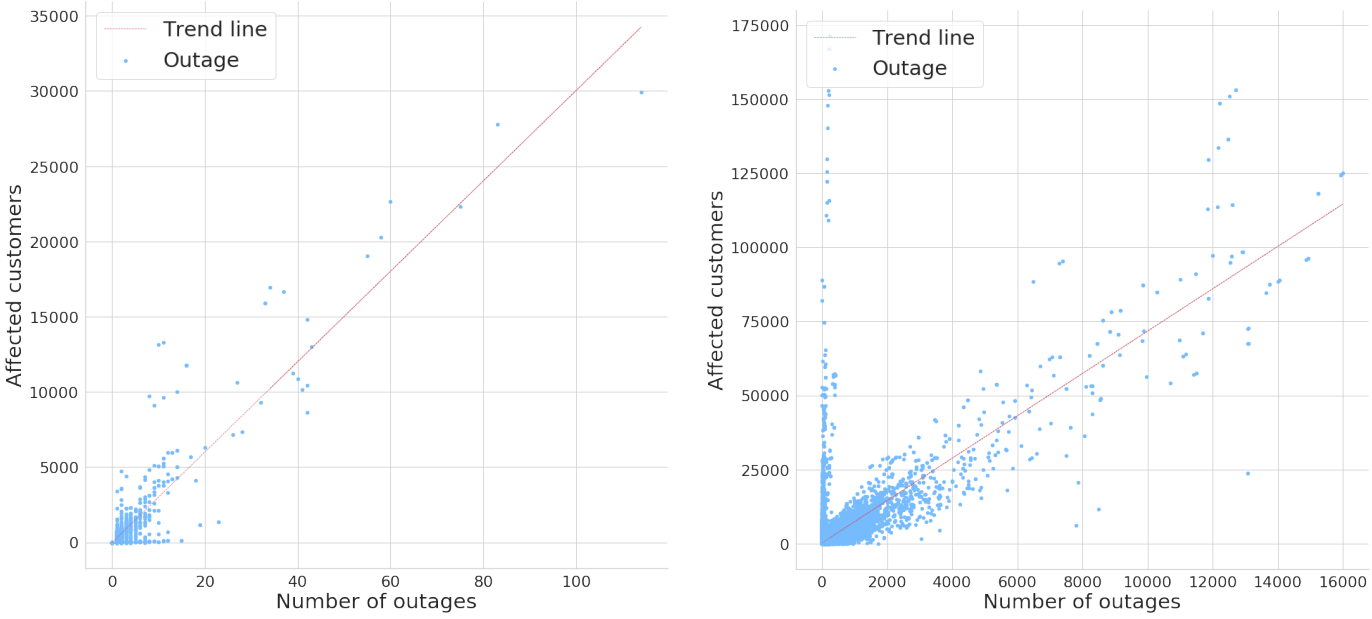
~~587~~ 17 587 storm objects have caused more than ~~10 outageswhich represent~~ ten outages, representing 12 percent of the whole dataset. ~~Based on the approximation, shown in Figure 5, one outage typically affects on 200 - 300 customers .~~ Figure 5 renders how many customers are typically affected by one outage. The figure contains all outages in both datasets, whether they are related to a storm or not. In the local dataset, usually 20-30 customers lose electricity in one outage. In the national dataset, only six customers usually lose electricity in one outage. We assume that this roots to different network topologies between the areas. Notably, in some rare cases, a much higher number of customers are affected. We assume that these cases occur typically in urban areas and are rare because the power network is mainly underground in these areas.

We use three classes designed together with power grid companies aiming at a simple "at glance" view for power grid operators. Class 0 represents no damage, class 1 low damage, and class 2 high damage. As the number of outages produced by a single storm object varies ~~a lot~~ significantly in the local and national datasets, we ~~end up defining~~ decided to define separate limits for the local and the national datasets. The detailed limits are listed in Table 2. Class 1 is defined such that it represents roughly 80 percent of all cases with at least one outage. Class sizes are highly imbalanced as most of the storm objects do not cause any damage.



(a) (b)

Figure 4. Number of storm objects per caused outages in (a) local dataset (b) national dataset.



(a) (b)

Figure 5. Relationship between number of outages and affected customers in (a) local dataset and (b) national dataset.

3.4 Classifying storm objects

We centered and normalized the data points by substracting the empirical mean and then dividing it by the empirical standard deviation. The hyperparameters were determined using random search 5-fold cross-validation (Bergstra and Bengio, 2012). To cope with the imbalanced class distribution, we generate artificial training samples using the synthetic minority over-sampling technique SMOTE (Chawla et al., 2002). The SMOTE creates new training samples based on their $k = 5$ nearest neighbors

Table 2. ~~Classes for local dataset~~Class definitions

Class	Outage limit in local ~~tdataset~~dataset	Local dataset size	Outage limit in national dataset	National dataset size
0	0	5 624	0	76 215
1	1-3	353	1- 140	14 417
2	≥ 4	181	≥ 141	3 085

following:

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i) \quad (1)$$

where x_i is an original class sample, x_{zi} is one of x_i's k nearest neighbor and λ is a random variable drawn uniformly from the interval $[0, 1]$. After augmentation, all classes have an equal number of samples, which reduces the tendency of classification methods to always predict the majority class.

Five different models were evaluated to classify storm objects. We omit the ~~more~~ mathematical definitions but shortly discuss the characteristics of different models and describe the implementation details chosen in this work.

Random forest classification (RFC) is based on a random ensemble of decision trees and ~~aggregate~~ aggregating results from individual trees to ~~final estimation~~the final estimate. Trees in the ensemble are constructed with four steps: 1) use bootstrapping to generate a random sample of the data, 2) randomly ~~selected~~ select a subset of features at each node, 3) determine the best split at the node using loss function, 4) grow the full tree (Breiman, 2001). RFC is ~~also~~ good to cope with high-dimensional data. It has also been found to provide adequate performance with imbalanced data (Tervo et al., 2019; Brown and Mues, 2012) ~~.~~and is widely used with weather data (e.g. Karthick et al. (2020); Cerrai et al. (2019); Lagerquist et al. (2017)). The method is prone to overfit, which is why hyperparameter-tuning is very important. Hyperparameters used in this work are listed in Table 3. We use RFC with the Gini impurity loss function. ~~Hyperparameters listed in Table 3.~~

Table 3. Hyperparameters for the RFC

Parameter	Value
Number of trees in the forest	500
Max depth	unlimited
Minimum nr. of samples to split	2
Minimun nr of samples to leaf	1
Features to consider for split	$\sqrt{\text{num. of feat.}}$
Max nro of leaf nodes	unlimited

Support Vector Classifiers (SVC) construct a hyper-plane or classification function ~~, in a high dimensional~~ in a high-dimensional feature space and maximize a distance between training samples and the hyperplane. The hyper-planes may be constructed with ~~non-linear~~ nonlinear kernels such as gaussian radial basis function (RBF) (Shawe-Taylor et al., 2004) that often reform a ~~non-linear~~ nonlinear classification problem to ~~linear~~a linear one. Operating in the high-dimensional feature space without additional computational complexity makes SVC an attractive choice to extract meaningful features from a ~~high dimensional data set. Furthermore, if~~ high-dimensional dataset. A domain-specific expert knowledge can also be capitalized on the kernel design. On the other hand, finding the correct kernel is often a difficult task. Training SVC is a convex optimization problem meaning that it has no local minima. Depending on the kernel, a training process may, however, be a very memory-intensive process.

Suppose the SVM output is assumed to be the log odds of a positive sample. In that case, one can fit a parametric model to obtain the posterior probability function and thus get probabilities for samples to belong to the particular class (Platt et al., 1999). For more details, we request the reader to consult for example Chang and Lin (2011) and Platt et al. (1999).

We implement the SVC in two phases. First, we separate class 0 (no outages) and other samples employing SVC with radial basis function (RBF), defined in Equation 2. Second, we distinguish classes 1 and 2 using SVC with a dot-product kernel defined in Equation 3 (Williams and Rasmussen, 2006). The second phase is performed only for the samples predicted to cause outages in the first phase. The approach is similar to the often-used one-vs-one classification, where a binary classifier is fitted for each pair of classes~~except that~~ . In our case different kernels were used for different pairs.

$$k_{RBF}(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma||\mathbf{x} - \mathbf{x}'||^2\right) \tag{2}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are two samples in the input space and γ is a kernel coefficient parameter.

$$k.(\mathbf{x}, \mathbf{x}') = \sigma_0 + x \cdot x' \tag{3}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are two samples in the input space and σ is a kernel inhomogenity parameter.

Gaussian Naive Bayes (GNB) (Chan et al., 1979) is a well-known and widely used method based on the Bayesian probability theory. The method assumes that all samples are independent and identically distributed (i.i.d), which does not naturally hold for the weather data. Despite the internal structure of the data, GNB is still ~~sometimes~~ used for weather ~~time series (for example Lindsay and Cox (2005)~~data (e.g. Kossin and Sitkowski (2009); Cintineo et al. (2014); Karthick et al. (2020)) and worth investigating ~~also~~ in this context. The classification rule in GNB is $\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y)$, where $P(y)$ is a frequency of class y and $P(x_i \mid y)$ is a likelihood of the ith feature assumed to be gaussian. Because of the naive i.i.d assumption, each likelihood can be estimated separately, which helps to cope with a curse of dimensionality and enable GNB to work relatively well with small datasets. On the other hand, estimating likelihoods can be done effectively and iteratively, enabling the GNB to scale to large datasets. As a downside, the simple method may lack expression power to perform well in a complex context.

Gaussian Processes (GP) (Rasmussen, 2003) is a non-parametric probabilistic method that interprets the observed data points as realizations of a Gaussian random process. GP is widely used for example in weather observation interpolation *kriging*

(Holdaway, 1996). ~~They are, however, computationally expensive and they tend to lose~~ GP is a very flexible and powerful but computationally expensive method, which tends to lose its power with high-dimensional data. GP ~~models hinge~~ hinges on a kernel function that encodes the covariance between different data points. As a kernel, we use a product of dot-product kernel (Equation 3) and pairwise kernel with laplacian distance (Rupp, 2015), defined in Equation 4. The kernel parameters were optimized on the training data by maximizing the log-marginal-likelihood.

$$k_{pairwise}(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma||\mathbf{x} - \mathbf{x}'||_1\right) \tag{4}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are two samples in the input space and γ is a kernel coefficient parameter.

Multilayer perceptrons (MLP) (Goodfellow et al., 2016) are the most basic form of ~~an artificial neural network~~artificial neural networks. Good results achieved by MLP in predicting storms (Ukkonen and Mäkelä, 2019), they are a natural choice to experiment ~~also~~ in this work. ~~The downsideof the method is~~ Neural networks are very adaptive methods as they can learn a representation of the input at their hidden layers. Unlike GNB, they do not make any assumptions about the distribution of the data. As a downside, MLP requires large amounts of data, and the training process is computing-intensive. They also have a large number of hyperparameters to be optimized, including the correct network topology.

We searched the correct model parameters and network topology for local and national datasets by running multiple iterations of random search 5-fold cross-validation to obtain the best possible micro average of F1-score (defined in Chapter 4) employing Talos library (Autonomio, 2020). ~~Final~~ The final setup composes of Nadam optimizer (Dozat, 2016), random normal initializer, and relu activation function for hidden layers. Binary cross-entropy was used as a loss function. Optimal network topology varied in different datasets~~.~~ : For the local dataset, the ~~used network contained~~ best results were obtained with a network containing three hidden layers with 75, 145, and 35 neurons. For the national dataset, the ~~network contained~~ best results were obtained with a network containing three hidden layers with 75, 195, and 300 neurons.

During the optimization process, the results varied between different setups from 0.6 to 0.95 in terms of F1-score.

4 Results

We used two different methods for splitting the data into training and test set. The first method is to use 25 percent of randomly picked samples in the test set. The second method is to construct a test set from a one-year continuous time range (2010-2011). Both approaches have their advantages. Continuous time range ensures that the model has not seen any autocorrelated samples caused by an internal structure of the weather data in the training phase (Roberts et al., 2017). However, having only ~~9~~ nine years of data from a relatively small geographical area, the continuous test set cannot contain many storms as most of the data needs to be reserved for the training process. Thus, the test set may only contain a single type of storms to which the model may work especially well or bad. Picking the test set randomly minimizes this risk and ~~provide more insight to~~ provides more insight into the model performance.

We evaluate the models with a weighted average of precision and recall, and both weighted and macro average of F1-score. Precision (Equation 5) reports how many samples are correctly predicted to belong to a class. Recall (Equation 6) tells how

many samples belonging to a class are found in the prediction. F1-score (Equations 7 and 8) calculates a harmonic mean of precision and recall. Finally, as the datasets are extremely imbalanced, we calculate a weighted average of the metrics utilizing a number of samples in each class and a macro average of F1-score using an average of F1-score of each class. A model with a higher macro average of F1-score performs better with small classes. The selected metrics do not take a distance between predicted and true class into account. It is naturally worse to predict, for example, class 0 (no damage) in the case of true class 2 (high damage) than in the case of true class 1 (low damage). We decided, however, to use metrics that measure the method performance properly with imbalanced classes.

$$Precision = \frac{1}{\sum_{c \in \mathcal{C}} |\hat{y}_c|} \sum_{c \in \mathcal{C}} \left(|\hat{y}_c| \frac{tp}{tp + fp} \right) \tag{5}$$

where $\mathcal{C}$ represents the set of classes, $\hat{y}$ predicted the class, tp true positives, and fp false positives.

$$Recall = \frac{1}{\sum_{c \in \mathcal{C}} |\hat{y}_c|} \sum_{c \in \mathcal{C}} \left(|\hat{y}_c| \frac{tp}{tp + fn} \right) \tag{6}$$

where $\mathcal{C}$ represents the set of classes, $\hat{y}$ predicted the class, tp true positives, and fn false negatives.

$$F1_{weighted} = \frac{1}{\sum_{c \in \mathcal{C}} |\hat{y}_c|} \sum_{c \in \mathcal{C}} \left(|\hat{y}_c| \frac{Precision_c \times Recall_c}{Precision_c + Recall_c} \right) \tag{7}$$

where $\mathcal{C}$ represents the set of classes, $\hat{y}$ predicted the class, Precision defined in Equation 5, and Recall defined in Equation 6.

$$F1_{macro} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left(\frac{Precision_c \times Recall_c}{Precision_c + Recall_c} \right) \tag{8}$$

where $\mathcal{C}$ represents the set of classes, Precision defined in Equation 5, and Recall defined in Equation 6.

Tables 4 and 5 divulge the results for each models using the local and national dataset respectively. Models trained with the local dataset can reach the better-weighted F1-score, while the best models trained with the national dataset provide a significantly better macro average of F1-score. The national dataset contains many more samples in classes 1 and 2~~which enable~~ , which enables models to learn the classes better and thus enhance the macro average of the F1-score. ~~Randomly chosen and continuous test set seems not~~ Whether the test set is randomly chosen or continuous does not seem to make a large difference in most cases. The only affected model is the RFC having contradictory better results trained with the continuous test set from the local dataset and the random test set from the national dataset. Assumingly, this squeal more about the unstable performance of RFC than the relevance of the dataset split method.

~~Confusion matrices are depicted in Figure 6. RFC provides the best results in terms of the selected metrics. Closer exploration revels, however, that this performance is largely due to the best performance in predicting class 0, which is the largest class. SVC results are one of the most balanced ones being the best only in the local dataset with a random test set but yielding good~~

stable results in all cases. The confusion matrix, shown in Figure 6b, displays that it is not the best model to predict class 0 but only a little share of true class 2 cases and the smallest share of true class 1 cases are predicted as class 0. That is to say, SVC misses the smallest number of harmful storms although it confuses in the amount of caused damage.

GP is another strong option that performs an even better job with class 0 while still providing good performance with class 2. A notable connecting aspect between GP and SVC is an almost identical kernel. Based on these experiments, in particular, RBF and pairwise kernels separate harmless and harmful samples from each other while dot-product kernel separates the classes 1 and 2 even better than exponential functions.

Using 15 threshold for detecting wind objects yields clearly better results than 20 threshold. For example SVC trained with national dataset using 20 threshold and randomly chosen test set provide only 0.48 macro average of F1-score being 12 percentage points below corresponding model using 15 threshold. 15 threshold have two major advantages compared to 20 . First, it provide significantly larger dataset and second, it is able to catch virtually all extratropical storms causing outages while 20 can not.

Table 4. Results for each models trained with the local dataset obtained from two local two power grid companies and (defined in Chapter 3.3)

Model	Split method	Precision	Recall	Weighted F1-score		Macro AVG F1-score	
		test	test	train	test	train	test
Random Forest Classifier (RFC)	Random	0.82	0.76	0.93	0.79	0.93	0.40
	Continuous	0.88	**0.91**	0.93	**0.89**	0.93	**0.48**
Support Vector Classifier (SVC)	Random	0.85	0.73	0.78	0.78	0.78	**0.44**
	Continuous	0.87	0.72	0.77	0.78	0.77	0.42
Gaussian Naive Bayes (GNB)	Random	**0.87**	0.61	0.59	0.70	0.59	0.42
	Continuous	**0.89**	0.59	0.59	0.69	0.59	0.40
Gaussian Processes (GP)	Random	0.84	0.70	1.0	0.76	1.0	0.43
	Continuous	0.85	0.67	0.94	0.74	0.94	0.41
Multilayer perceptor (MLP)	Random	0.82	**0.81**	0.98	**0.80**	0.91	0.41
	Continuous	0.81	0.79	0.97	0.80	0.91	0.41

The confusion matrices are depicted in Figure 6. RFC provides the best results in terms of the selected metrics. However, closer exploration reveals that this performance is largely due to the best performance in predicting class 0, which is the largest class. SVC results are one of the most balanced ones being the best only in the local dataset with a random test set but yielding good stable results in all cases. The confusion matrix, shown in Figure 6b, displays that it is not the best model to predict class

Table 5. Results for each models trained with the national dataset covering whole Finland ~~and~~ (defined in Chapter 3.3)

Model	test set split method	Precision	Recall	Weighted F1-score		Macro AVG F1-score	
		test	test	train	test	train	test
Random Forest Classifier (RFC)	Random	**0.83**	**0.84**	1.0	**0.83**	1.0	**0.62**
	Continuous	**0.77**	**0.81**	1.0	**0.78**	1.0	0.40
Support Vector Classifier (SVC)	Random	0.81	0.61	0.68	0.68	0.68	0.60
	Continuous	0.62	0.60	0.60	0.60	0.60	0.60
Gaussian Naive Bayes (GNB)	Random	0.75	0.60	0.66	0.66	0.45	0.39
	Continuous	**0.77**	0.60	0.45	0.66	0.45	0.40
Gaussian Processes (GP)	Random	0.57	0.56	0.71	0.55	0.71	0.55
	Continuous	0.67	0.65	0.94	0.65	0.94	**0.61**
Multilayer perceptor (MLP)	Random	0.79	0.75	0.94	0.77	0.90	0.52
	Continuous	0.76	0.78	0.93	0.78	0.85	0.40

0, but only a little share of true class 2 cases and the smallest share of true class 1 cases are predicted as class 0. That is to say, SVC misses the smallest number of destructive storms, although it confuses in the amount of caused damage.

GP is another strong option that performs even better with class 0 while still providing good performance with class 2. A significant connecting aspect between GP and SVC is an almost identical kernel. Based on these experiments, RBF and pairwise kernels separate harmless and harmful samples from each other while dot-product kernel separates the classes 1 and 2 even better than exponential functions. We select GP for further analysis in this paper since it provides the best performance in class 2.

Using the 15 $\mathrm{m\,s^{-1}}$ threshold for detecting storm objects yields clearly better results than the 20 $\mathrm{m\,s^{-1}}$ threshold. For example, SVC trained with the national dataset using the 20 $\mathrm{m\,s^{-1}}$ threshold and randomly chosen test set provide only 0.48 macro average of F1-score being 12 percentage points below corresponding model using the 15 $\mathrm{m\,s^{-1}}$ threshold. The 15 $\mathrm{m\,s^{-1}}$ threshold have two major advantages compared to the 20 $\mathrm{m\,s^{-1}}$. First, it provides a significantly larger dataset and second, in contrast to the 20 $\mathrm{m\,s^{-1}}$ threshold, it is able to catch virtually all extratropical storms causing outages.

4.1 Feature importances in the model performance

The relevance of the individual predictive features can be explored by using the permutation test, as done by Breiman (2001) . First, the baseline score of the fitted model is calculated using the test set. Then each feature is randomly permuted, and the difference in the scoring function is calculated. The random permutation is repeated 30 times for each parameter, and the average of the results is used. The procedure offers information on how important the feature is to obtain good results. It should

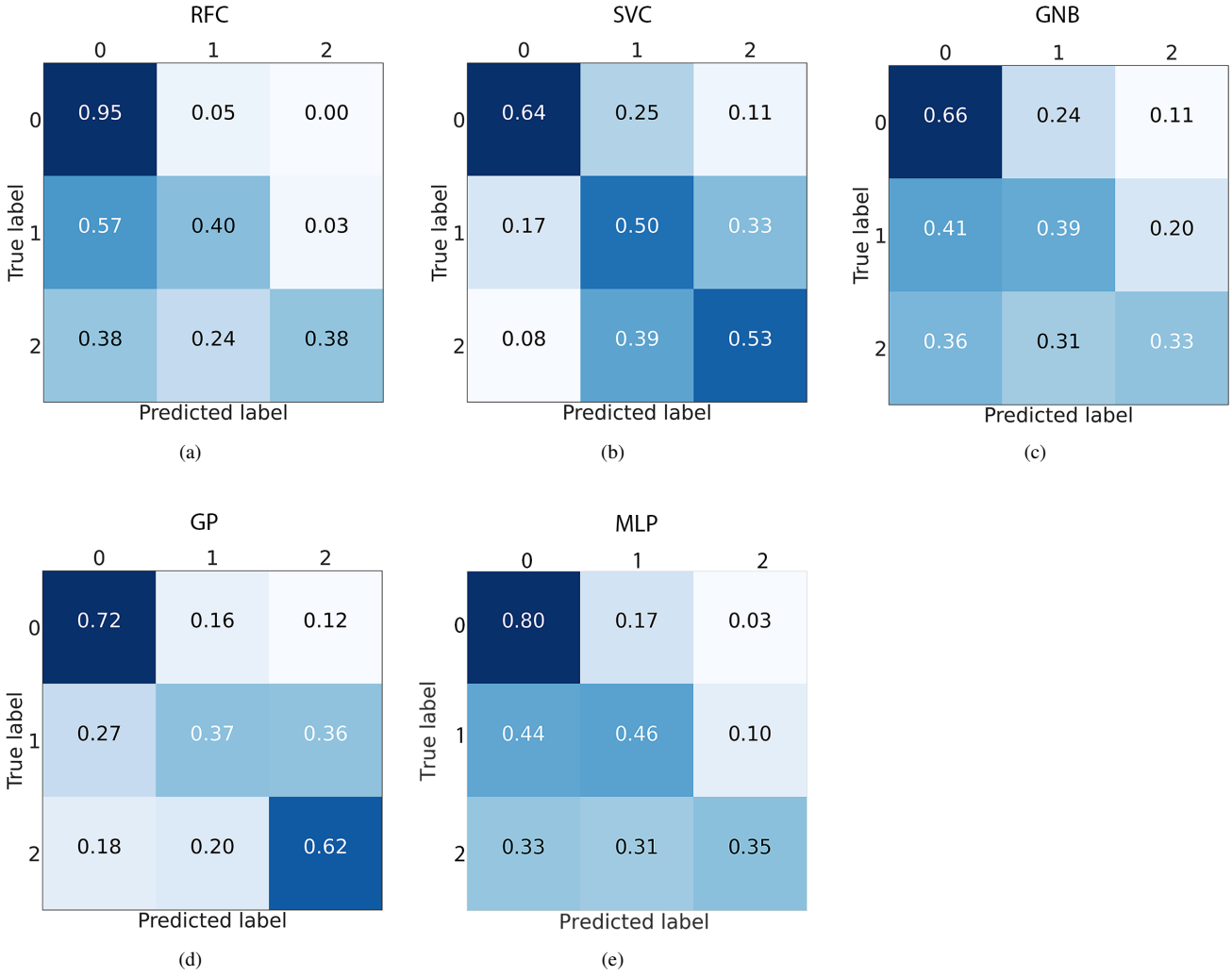


Figure 6. Confusion matrices produced using the randomly selected national dataset and (a) RFC (b) SVC (c) GNB (d) GP (e) MLP. Each cell of the confusion matrices represents a share of predictions having a corresponding combination of predicted and true class. For example, the middle right cell tells the share of samples belonging to class 1 but predicted to have class 2.

be mentioned that highly correlated features may get low importance as other features work as a proxy to the permuted feature. However, using completely independent features is not possible in weather data since weather parameters are often dependent on each other, and eliminating even the most apparent pairs from the used features impaired the results in our experiments.

We used the macro average of F1 defined in Equation 8 as a scoring function and the randomly selected test set from the national data. The relevance is shown in Figure 7. Most features show at least little relevance for the results. The first twelve features are significantly more relevant than the rest. The most important features contain at least one representative of all

meteorological parameters used in the training. In other words, all employed meteorological parameters are important for the prediction, while different aggregations are contributing to the "fine-tuning" of the model.

As Figure 7 shows, the most significant parameter regarding our model performance is the average wind speed. Numerous studies support our result of wind being the most important damaging factor (Virot et al., 2016; Valta et al., 2019; Jokinen et al., 2015). The studies are, however, highlighting the importance of maximum wind gusts instead of the average wind. Surprisingly, in our analysis, the wind gust speed does not belong to the most critical parameters. Instead, maximum mixed layer height, related to the wind gustiness, contributes crucially to the model performance. The dependencies between predictive features might be one reason for some parameters to have a lower rank in the results.

The stand mean diameter and height are the most important features regarding the forest parameters, which corresponds to our expectations. Previous studies also state these features to influence the wind damage in forests (Pellikka and Järvenpää, 2003) and hence indirectly electricity grids. As Pellikka and Järvenpää (2003) and Suvanto et al. (2016) discuss, also the age of the forest has an impact on storm damages. However, in the feature importance test, forest age does not seem to contribute significantly to the prediction outcome.

The most important object feature is the size of the object. Object movement speed and direction did not contribute strongly to the results. However, previous studies indicate that besides the size of the impacted area, the duration of strong winds – i.e., the propagation speed of the system – influences also the amount of damage (Lamb and Knud, 1991).

4.2 Case Examples

We illustrate the prediction produced using GP classification method with the three most interesting examples of well-known storms in Figure 8a. We chose the cases among a number of test cases to ~~represent~~ illustrate the strengths and weaknesses of the method. The examples are chosen from the randomly picked test set, which was not used to train the model. Because of the random sample, we cannot represent the entire prediction of individual storms, only individually picked time steps. In two of the example cases, the model performs well (storms Tapani and Pauliina) and in one (storm Rauli) with less accurate prediction results.

4.2.1 Event 1: Extratropical Storm Tapani (26 December 2011)

The first example is one of the most known extratropical storms in Finland. Storm Tapani, ~~known also~~ also known as Cyclone Dagmar (Kufeoglu and Lehtonen, 2015), was a rare winter storm, causing ~~wide~~ broad and long-lasting electricity interruptions. Extreme wind gusts of over 30 $\mathrm{m\,s^{-1}}$ caused widespread damage, especially in the southern and western parts of the country. Approximately 570 000 households were left without electricity, causing 30 million euros repair costs and 80 million euros of monetary compensation for electricity distribution companies to their customers ~~. (Hanninen and Naukkarinen, 2012)~~ (Hanninen and Naukkarinen, 2012). Exceptionally warm December and the Boxing day being the warmest in 50 years (Finnish Meteorological Institute, 2011) resulted in wet and unfrozen soil~~, thus, .~~ Thus, the trees were poorly anchored and exposed to ~~major~~ significant storm damage.

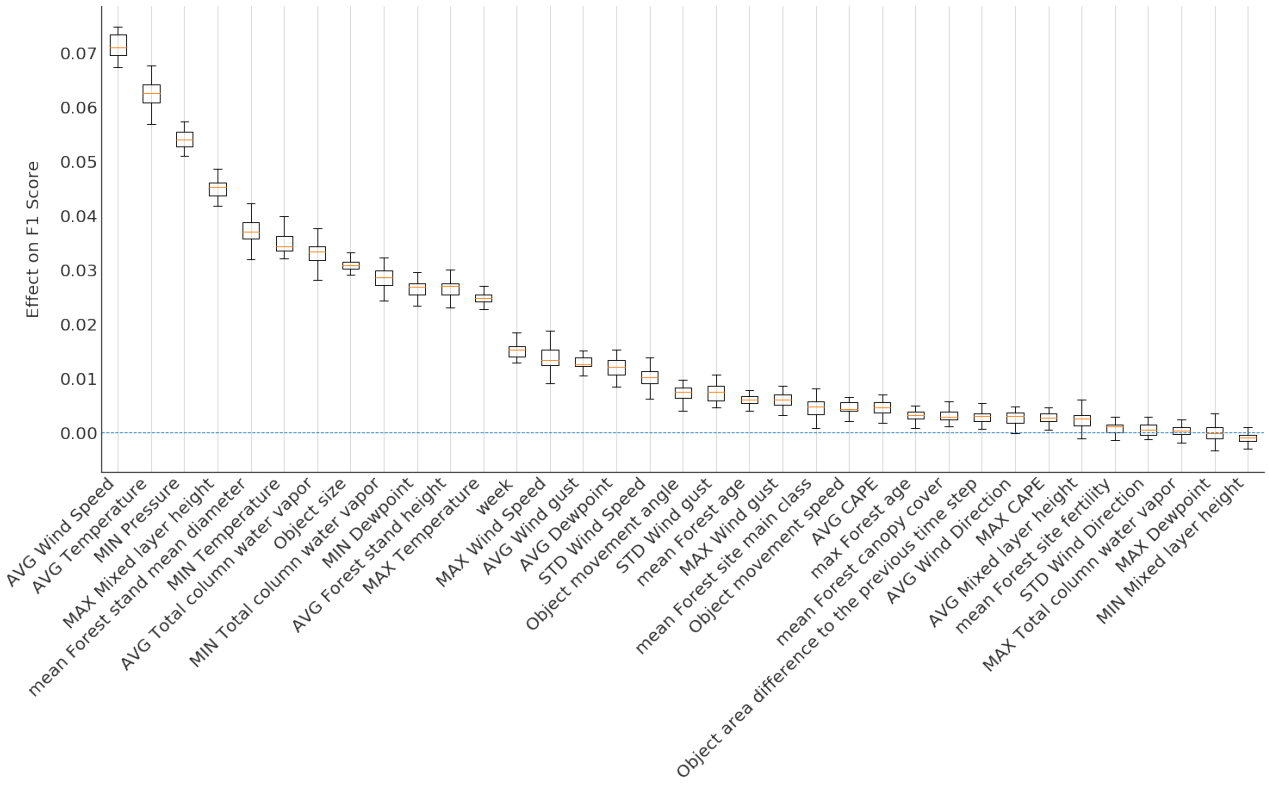


Figure 7. Permutation feature importance using the GP classification method trained with the randomly selected national dataset. The higher effect on the F1 score is (y-axis), the bigger is the significance.

Figure 8a represents the outage prediction (raster-covered areas) and the actual, true classes (numbers) based on the damage data at 15:00 UTC, 26 December 2011. Wide areas in central and western parts of Finland are predicted to have *high* ~~,~~ (class 2) damages. The predicted class is in line with the true class. Also, the damage areas of the storm correlate with the wind gust observations of the Finnish Meteorological Institute. The strongest gusts ~~situated~~ occurred in western (15-27 $\mathrm{m\,s^{-1}}$) and southern (18-28 $\mathrm{m\,s^{-1}}$) Finland and north-western part of Lapland (13-31$\mathrm{m\,s^{-1}}$) (Finnish Meteorological Institute, 2020). In the rest of Finland, the maximum wind gusts remained between 10-15 $\mathrm{m\,s^{-1}}$, and therefore the damages were ~~also~~ minor. Overall, the model ~~works~~ predicted the damages accurately in this particular example~~accurately.~~.

4.2.2 Event 2: Extratropical Storm Rauli (27 August 2016)

Extratropical storm Rauli was ~~exceptionally strong for the summer season~~an exceptionally strong summer storm, especially regarding the impacts. It caused severe damages ~~for~~ to the power grid in the western and middle parts of Finland for various reasons. The trees ~~had leaveson~~were carrying leaves, the soil was wet after a rainy August, the strong wind areas of Rauli were widely spread, and the solar radiation was intensifying the wind gusts during the afternoon (Finnish Meteorological Institute,

2016). Rauli was impacting especially the middle and southern parts of Finland, which are also the most densely populated areas. The power outages were increasing rapidly in the middle part of Finland, starting at midday and reaching the highest values, 200 000 households without electricity (Ilta-Sanomat, 2016), around 5 pm. The winds were blowing exceptionally long, nearly 24 hours. The typical duration of summer storms is between 6-12 hours.

Figure 8b shows the predicted outages and true classes at 12:00 UTC, 27 August 2016. In this particular time step, the model is over-predicting the class, however, the predicted outage area seems to correlate with the wind gust ~~maxims during~~ maximums of that afternoon. The strongest wind gusts were measured in the southern and middle parts of the country, maximum gusts reaching on land stations up to 24,9 $\mathrm{m\ s^{-1}}$ (Klemettilä, Vaasa and Maaninka, Pohjois-Savo) and on wide areas up to 20 $\mathrm{m\ s^{-1}}$ apart from the northern part of Finland.

4.2.3 Event 3: Extratropical storm Pauliina (22 June 2018)

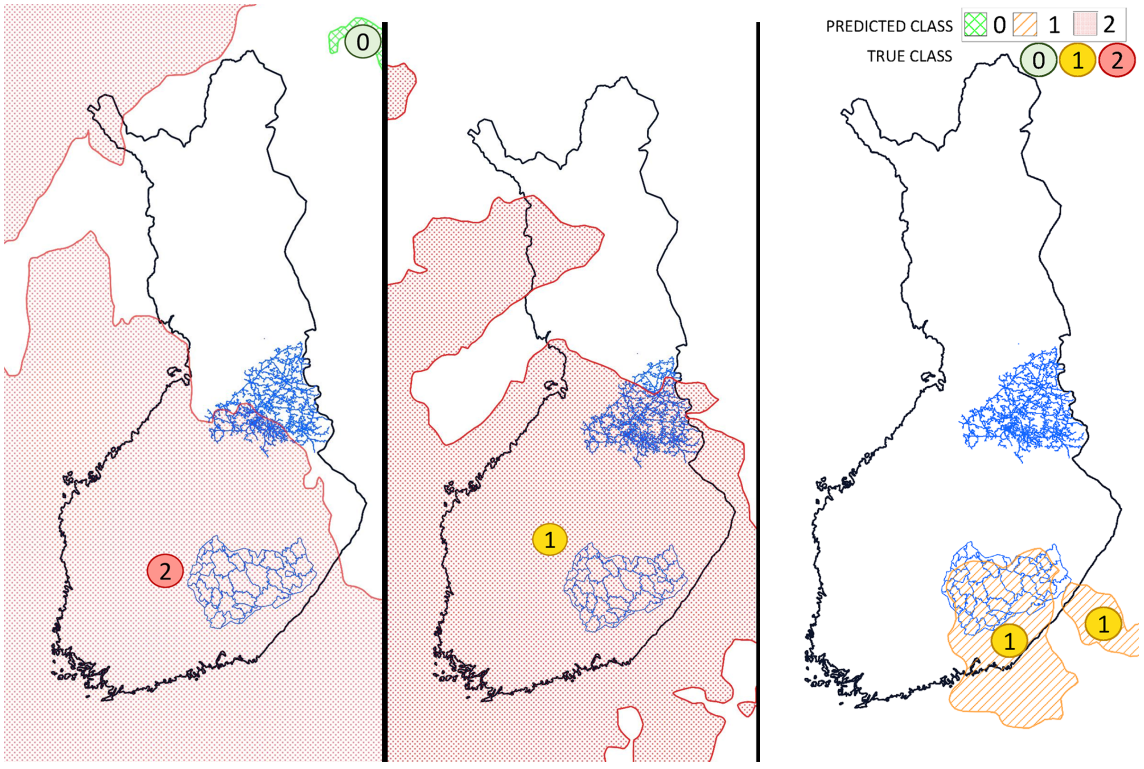
The last example is a strong extratropical storm, called Pauliina ~~which (Finnish Meteorological Institute, 2018)~~ (Finnish Meteorological Ins that caused numerous power outages in Finland. The most significant part of the power outages happened in the network of power grid company JSE ~~which is~~ included in the local dataset. The highest peak in the damages was reached between 6 and 8 ~~p.m~~ pm with over 28 000 households without electricity. The strongest wind gust on land reached 22,7 $\mathrm{m\ s^{-1}}$ ~~on~~ in Helsinki, Kumpula observation station, and the inland gusts were widely between 15-20 $\mathrm{m\ s^{-1}}$ (Finnish Meteorological Institute, 2020; Finnish Meteorological Institute, Twitter). The strong wind gusts continued until the dawn of the 23rd of June.

Figure 8c presents the predicted and true damage classes at 01:00, UTC, 22 June 2018. We chose extratropical storm Pauliina as an example storm for two reasons: 1) Pauliina represents a *low damage* class 2) Pauliina represents a rare, summer-season extratropical storm. Figure 8c shows the predicted and true classes correlating. While weather warnings were ~~given~~ issued to large areas in southern and middle parts of Finland, (myrskyvaroitus.com, 2018) predicted and true damage to the power grid occurred in a relatively small geographical area. ~~This example shows the potential added value of the model compared to weather warnings providing more accurate information to the power grid operators.~~

5 Discussion and conclusions

This paper ~~introduced~~ introduces a novel method to predict the damage potential of extratropical storms to power grids. The method consists of identifying ~~wind~~ storm objects by contouring surface wind gust ~~field with~~ fields with the 15 $\mathrm{m\ s^{-1}}$ threshold along with pressure objects with a 1000 hPa threshold, tracking the objects, and ~~classifying them in~~ then classifying them into three classes based on their damage potential to the power grid. For the classification task, we evaluated five different machine learning methods, all employing in a total of 35 predictive features and trained with ~~8~~ eight years of power outage data from Finland.

~~The most balanced results were gained with the Support Vector Classifier~~Both Gaussian Processes and Support Vector Classifiers provided good results. The model recognizes harmful storm ~~cells~~ objects well and can distinguish extremely harmful ~~cells~~ objects among others adequately. While the results still ~~left~~ leave a lot to improve, the developed model can be already



(a) Tapani, 26 December 2011 11:00 (b) Rauli, 27 August 2016 10:00 (c) Pauliina, 22 June 2018 01:00

Figure 8. Selected examples (a) Extratropical storm Tapani (b) Extratropical storm Rauli (c) Extratropical storm Pauliina, produced by employing the SVC model trained with the national dataset. The storm objects are ~~coloured~~ colored based on the predicted class while the true class is stated as a ~~coloured~~ colored number over the object. The time is represented as UTC time.

used to ~~help decisions made~~ support decisions in power grid companies. ~~The~~ In some cases, the model is able to provide a more specific and geospatially accurate prediction of ~~caused~~ potential damage to the power grid than~~for example~~ , for example, weather warning. The evaluation was, however, based on the ERA5 reanalysis data. Using the method in an operational setting would require weather prediction data, which introduces additional uncertainty to the outage prediction.

~~The work~~ The presented object-based approach has both advantages and disadvantages. Extracting storm objects in advance preprocesses the data for machine-learning techniques, such as RFC, which do not perform feature learning. It enables machine-learning methods to focus only on the relevant parts of the data. Methods not containing feature learning, such as RFC and logistic regression, have been found to outperform neural networks for forest (Hart et al., 2019) and weather data (Tervo et al., 2019). It also leads to significantly faster training times. Processing objects instead of the grid makes it also easier to track and use object attributes such as age, speed, and movement. Moreover, objects are easy to visualize, and user interfaces may be enriched with related actions such as tracking and alarms.

On the other hand, storm objects use only aggregated attributes, which may decrease the classification accuracy when predictive features vary significantly under the storm object area. Several machine-learning methods, i.e., deep neural networks, could be trained to employ those local features to gain better accuracy. Such methods could also utilize three-dimensional data.

Extracting storm objects requires a fixed threshold of wind gust and pressure, which may vary depending on the characteristics of geospatial locations. Nevertheless, the previous studies indicate the critical threshold to wind gust speed to be the same for almost entire geospatial domain of this work (Gardiner et al., 2013). Moreover, the correct threshold may vary depending on the data source. When extending the geospatial domain or changing the data source, this might become a more serious issue, and different thresholds might be needed. One possibility to determine the optimal threshold might be to use specific quantiles of the parameter values, but this would need further investigation.

The work opens several possible avenues for further studies. ~~Soil~~ It would be interesting to compare the current solution with a grid-based approach and deep neural networks. Including data on soil moisture, soil temperature, and leaf index would most ~~probably~~ likely enhance the results~~as~~ , if available with sufficient spatial and temporal resolution, since they would provide critical information about the environmental conditions. Different thresholds could be investigated as well, especially for pressure objects where lower thresholds might yield better results. By design, applying the method ~~on other regions should be possibleas well by using available~~ to other regions is possible, but it is subject to the availability of power outage records, forest inventory, impact and meteorological data. For the classification task, carefully designed Bayesian networks could provide good results as well. Especially in the randomly selected test set, data may be autocorrelated, which may lead to unrealistically good results. We have addressed this issue by also using a continuous time series (from 2010 to 2011) for the test set. The evaluation could also be extended with a leave-one-day-out or leave-one-week-out method where for each week one day or for each month one week is hold out for validation purposes.

End users, especially expert users like duty forecasters, might benefit from the uncertainty information originating as the probabilistic prediction of the classification model. However, the presentation of such information should be very carefully chosen not to mislead non-expert users for overconfidence.

Experiments in this study were conducted with ERA5 reanalysis and additional forest data. As the method employs common features existing also in various other datasets, data provided by other vendors could be used as well. ~~In the future~~By employing weather forecasts as input, this method could be used as a base for a decision support tool and as a part of an existing early warning system ~~, for both ,~~ for both duty forecasters of national hydro-meteorological centers ~~as well as~~ and operators of electricity transmission companies.

Code and data availability. The source code is available in the repositories https://github.com/fmidev/sasse-era5-smartmet-grid and https://github.com/fmidev/sasse-polygon-process. ERA5 data may be downloaded from the Copernicus Climate Data Store: https://cds.climate.copernicus.eu. Forest inventory may be downloaded from LUKE open data service: http://kartta.luke.fi/index-en.html. The power outage data is propriety data which the authors have no property rights to distribute.

Appendix A: Gaussian distribution fitted to the storm object features

A1 ~~Local dataset~~

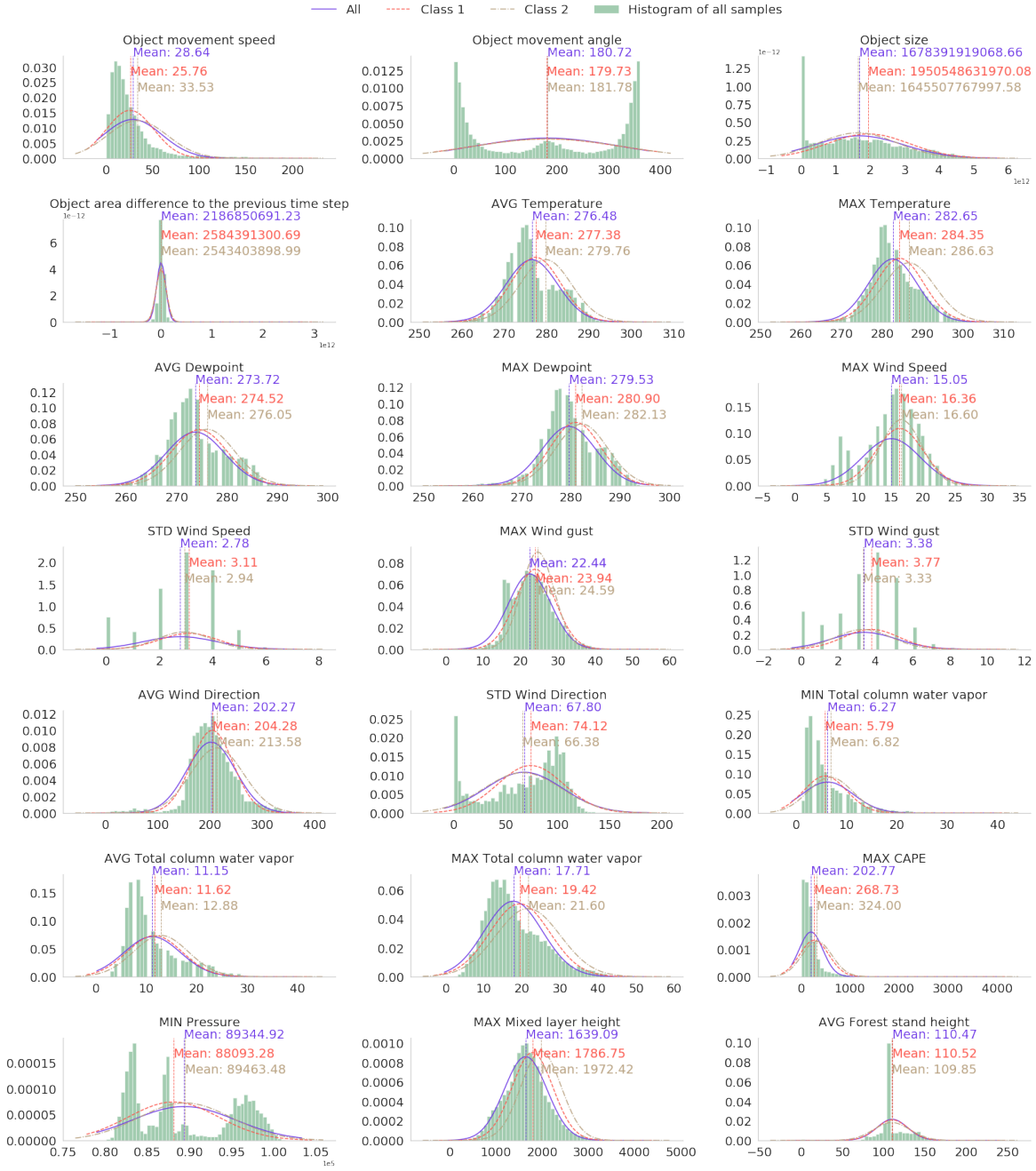


Figure A1. ~~Distribtution~~ Histogram of ~~the~~ and fitted Gaussian distribution of selected predictive parameters ~~of~~ in **the local dataset**. The Gaussian distribution is fitted separately to all ~~storm objects (~~samples ~~)~~ and ~~objects~~ samples with ~~class 2. Local dataset~~little outages and many outages (classes 1 and 2 specified in Section 3.3).

A1 ~~National dataset~~

Author contributions. RT conceptualized, designed, and developed the method. IL contributed with meteorological expertise, such as selecting used data and meteorological features along with correct thresholds. IL also helped in analyzing the performance. AM provided supervision from a meteorological perspective and AJ from a machine learning perspective. All contributed in presenting the results.

Competing interests. The authors declare that they have no conflict of interests.

Acknowledgements. The authors ~~thanks project partners~~ express their gratitude to Järvi-Suomen Energia, Loiste Sähköverkko, and Imatran Seudun Sähkönsiirto for ~~data and~~ sharing data and their experience.

References

Allen, M., Fernandez, S., Omitaomu, O., and Walker, K.: Application of hybrid geo-spatially granular fragility curves to improve power outage predictions, Journal of Geography & Natural Disasters, 4, 1–6, https://doi.org/10.4172/2167-0587.1000127, 2014.

Autonomio: Talos (software), http://github.com/autonomio/talos., 2020.

Barredo, J.: No upward trend in normalised windstorm losses in Europe: 1970-2008, Natural Hazards and Earth System Sciences, 10, https://doi.org/10.5194/nhess-10-97-2010, 2010.

Bergstra, J. and Bengio, Y.: Random search for hyper-parameter optimization, Journal of Machine Learning Research, 13, 281–305, 2012.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Brown, I. and Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets, Expert Systems with Applications, 39, 3446–3453, 2012.

Campbell, R. J. and Lowry, S.: Weather-related power outages and electric system resiliency, Tech. rep., Congressional Research Service, Library of Congress Washington, DC, 2012.

Cerrai, D., Wanik, D. W., Bhuiyan, M. A. E., Zhang, X., Yang, J., Frediani, M. E., and Anagnostou, E. N.: Predicting storm outages through new representations of weather and vegetation, IEEE Access, 7, 29 639–29 654, 2019.

Chan, T., Golub, G., and LeVeque, R.: Updating formulae and a pairwise algorithm for variances computing sample, in: COMPSTAT, p. 30, Springer Science & Business Media, 1979.

Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, ACM transactions on intelligent systems and technology (TIST), 2, 27, 2011.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, 321–357, 2002.

Chen, P.-C. and Kezunovic, M.: Fuzzy logic approach to predictive risk analysis in distribution outage management, IEEE Transactions on Smart Grid, 7, 2827–2836, 2016.

Chen, S. H., Jakeman, A. J., and Norton, J. P.: Artificial intelligence techniques: an introduction to their use for modelling environmental systems, Mathematics and computers in simulation, 78, 379–400, 2008.

Cintineo, J. L., Pavolonis, M. J., Sieglaff, J. M., and Lindsey, D. T.: An empirical model for assessing the severe weather potential of developing convection, Weather and Forecasting, 29, 639–653, 2014.

Csilléry, K., Kunstler, G., Courbaud, B., Allard, D., Lassegues, P., Haslinger, K., and Gardiner, B.: Coupled effects of wind-storms and drought on tree mortality across 115 forest stands from the Western Alps and the Jura mountains, Global change biology, 23, https://doi.org/10.1111/gcb.13773, 2017.

Donat, M. G., Leckebusch, G. C., Wild, S., and Ulbrich, U.: Future changes in European winter storm losses and extreme wind speeds inferred from GCM and RCM multi-model simulations, Natural Hazards and Earth System Sciences, 11, 1351–1370, https://doi.org/10.5194/nhess-11-1351-2011, https://www.nat-hazards-earth-syst-sci.net/11/1351/2011/, 2011.

Dozat, T.: Incorporating nesterov momentum into adam, 2016.

Eskandarpour, R. and Khodaei, A.: Machine Learning Based Power Grid Outage Prediction in Response to Extreme Events, IEEE Transactions on Power Systems, 32, 3315–3316, 2017.

European Centre for Medium-Range Weather Forecasts: ERA5 Reanalysis, https://doi.org/10.5065/D6X34W69, 2017.

European Centre for Medium-Range Weather Forecasts: ERA5: large 10m winds, https://confluence.ecmwf.int/display/CKB/ERA5:+large+10m+winds, 2019.

Finnish Energy, E. O.: Energy distribution interruptions 2010-2018 - Finnish Energy, https://energia.fi/julkaisut/materiaalipankki/sahkon_keskeytystilastot_2010-2018.html#material-view, 2010-2018.

Finnish Energy, E. O.: Energy distribution interruptions 2011 - Finnish Energy, Tech. rep., Finnish Energy ry, https://energia.fi/files/605/Keskeytystilasto_2011.pdf, 2011.

Finnish Energy, E. O.: Energy distribution interruptions 2013 - Finnish Energy, Tech. rep., Finnish Energy ry, https://energia.fi/files/605/Keskeytystilasto_2013.pdf, 2013.

Finnish Meteorological Institute: Climate Statistics Report (December 2011), pp. 1–24, 2011.

Finnish Meteorological Institute: Rauli nousi myrskyjen raskaaseen sarjaan | Ilmastokatsaus – Ilmatieteen laitos, Ilmastokatsaus, http://www.ilmastokatsaus.fi/2016/08/29/rauli-nousi-myrskyjen-raskaaseen-sarjaan/, 2016.

Finnish Meteorological Institute: Kesäkuun 2018 kuukausikatsaus | Ilmastokatsaus – Ilmatieteen laitos, http://www.ilmastokatsaus.fi/2018/07/04/kesakuun-2018-kuukausikatsaus/, 2018.

Finnish Meteorological Institute: Open data, observation download service, https://en.ilmatieteenlaitos.fi/download-observations, 2020.

Finnish Meteorological Institute, Twitter: Ilmatieteen laitos Twitterissä: "#Juhannus'aaton-#myrsky #Pauliina lukuina klo 15.30: kovin puuska 32,5 m/s ja 10min keskituuli 26,5 m/s (Kaskinen Sälgrund), maa-alueilla Helsinki Kumpula puuska 22,7 m/s. Vaasa Klemettilä 24h sademäärä 41 mm." / Twitter, https://twitter.com/meteorologit/status/1010144460703969280, 2020.

Gardiner, B., Blennow, K., Carnus, J., Fleischner, P., Ingemarson, F., Landmann, G., Lindner, M., Marzano, M., Nicoll, B., Orazio, C., Peyron, J., Reviron, M., Schelhaas, M., Schuck, A., Spielmann, M., and Usbeck, T.: Destructive storms in European Forests: Past and Forthcoming Impacts. Final report to European Commission - DG Environment,, European Forest Institute, 2010.

Gardiner, B., Schuck, A., Schelhaas, M.-J., Orazio, C., Blennow, K., and Nicoll, B.: Living with Storm Damage to Forests What Science Can Tell Us What Science Can Tell Us, European Forest Institute, EFI, Joensuu, Finland, https://doi.org/10.13140/2.1.1730.2400, 2013.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, in: Deep Learning, pp. 164–223, MIT Press, 2016.

Govorushko, S. M.: Natural processes and human impacts: Interactions between humanity and the environment, Springer Science & Business Media, 2011.

Gregow, H., Laaksonen, A., and Alper, M. E.: Increasing large scale windstorm damage in Western, Central and Northern European forests, 1951-2010, Scientific Reports, 7, https://doi.org/10.1038/srep46397, 2017.

Guikema, S. D., Quiring, S. M., and Han, S. R.: Prestorm Estimation of Hurricane Damage to Electric Power Distribution Systems, Risk Analysis, 30, 1744–1752, https://doi.org/10.1111/j.1539-6924.2010.01510.x, 2010.

Guikema, S. D., Nateghi, R., Quiring, S. M., Staid, A., Reilly, A. C., and Gao, M.: Predicting Hurricane Power Outages to Support Storm Response Planning, IEEE Access, 2, 1364–1373, https://doi.org/10.1109/ACCESS.2014.2365716, 2014.

Haarsma, R. J., Hazeleger, W., Severijns, C., De Vries, H., Sterl, A., Bintanja, R., Van Oldenborgh, G. J., and Van Den Brink, H. W.: More hurricanes to hit western Europe due to global warming, Geophysical Research Letters, 40, 1783–1788, https://doi.org/10.1002/grl.50360, 2013.

Han, S. R., Guikema, S. D., and Quiring, S. M.: Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models, Risk Analysis, 29, 1443–1453, https://doi.org/10.1111/j.1539-6924.2009.01280.x, 2009.

Hanewinkel, M.: Neural networks for assessing the risk of windthrow on the forest division level: a case study in southwest Germany, European Journal of Forest Research, 124, 243–249, 2005.

Hanninen, K. and Naukkarinen, J.: 570 000 customers experienced power losses at the end of the year, (Loppuvuoden sähkökatkoista kärsi 570 000 asiakasta - Energiateollisuus, http://www.mynewsdesk.com/fi/pressreleases/loppuvuoden-saehkoekatkoista-kaersi-570-000-asiakasta-725038, 2012.

Hart, E., Sim, K., Kamimura, K., Meredieu, C., Guyon, D., and Gardiner, B.: Use of machine learning techniques to model wind damage to forests, Agricultural and forest meteorology, 265, 16–29, 2019.

He, J., Wanik, D. W., Hartman, B. M., Anagnostou, E. N., Astitha, M., and Frediani, M. E.: Nonparametric Tree-Based Predictive Modeling of Storm Outages on an Electric Distribution Network, Risk Analysis, 37, 441–458, https://doi.org/10.1111/risa.12652, 2017.

Hersbach, H., Bell, W., Berrisford, P., Horányi, A., J., M.-S., Nicolas, J., Radu, R., Schepers, D., Simmons, A., Soci, C., and Dee, D.: Global reanalysis: goodbye ERA-Interim, hello ERA5, ECMWF newsletter, pp. 17–24, https://doi.org/10.21957/vf291hehd7, https://www.ecmwf.int/node/19027, 2019.

Holdaway, M. R.: Spatial modeling and interpolation of monthly temperature using kriging, Climate Research, 6, 215–225, 1996.

Ilta-Sanomat: Rauli-myrsky repii Suomea: lähes 200000 taloutta ilman sähköjä - Kotimaa - Ilta-Sanomat, https://www.is.fi/kotimaa/art-2000001248969.html, 2016.

IPCC: Impacts of 1.5 C global warming on natural and human systems, Global warming of 1.5° C. An IPCC Special Report, 2018.

Jokinen, P., Vajda, A., and Gregow, H.: The benefits of emergency rescue and reanalysis data in decadal storm damage assessment studies, Advances in Science and Research, 12, 97–101, https://doi.org/10.5194/asr-12-97-2015, 2015.

Kankanala, P., Pahwa, A., and Das, S.: Regression models for outages due to wind and lightning on overhead distribution feeders, in: Power and Energy Society General Meeting, 2011 IEEE, pp. 1–4, IEEE, 2011.

Kankanala, P., Pahwa, A., and Das, S.: Estimation of Overhead Distribution System Outages Caused by Wind and Lightning Using an Artificial Neural Network, in: International Conference on Power System Operation & Planning, 2012.

Kankanala, P., Das, S., and Pahwa, A.: AdaBoost$^+$: An Ensemble Learning Approach for Estimating Weather-Related Outages in Distribution Systems, IEEE Transactions on Power Systems, 29, 359–367, 2014.

Karthick, S., Malathi, D., Sudarsan, J., and Nithiyanantham, S.: Performance, evaluation and prediction of weather and cyclone categorization using various algorithms, Modeling Earth Systems and Environment, pp. 1–9, 2020.

Kossin, J. P. and Sitkowski, M.: An objective model for identifying secondary eyewall formation in hurricanes, Monthly Weather Review, 137, 876–892, 2009.

Kron W., Schuck A.: After the floods, Tech. rep., Münchener Rückversicherungs-Gesellschaft (Munich Reinsurance Company), https://www.munichre.com/content/dam/munichre/global/content-pieces/documents/302-08121_en.pdf/_jcr_content/renditions/original./302-08121_en.pdf, 2013.

Kufeoglu, S. and Lehtonen, M.: Cyclone Dagmar of 2011 and its impacts in Finland, in: IEEE PES Innovative Smart Grid Technologies Conference Europe, vol. 2015-January, IEEE Computer Society, https://doi.org/10.1109/ISGTEurope.2014.7028868, 2015.

Lagerquist, R., McGovern, A., and Smith, T.: Machine learning for real-time prediction of damaging straight-line convective wind, Weather and Forecasting, 32, 2175–2193, 2017.

Lamb, H. and Knud, F.: Historic Storms of the North Sea, British Isles and Northwest Europe - Google Books, https://books.google.fi/books?hl=en&lr=&id=P4n1z9rOh5MC&oi=fnd&pg=PR8&ots=ewtYPSlQ1H&sig=VJJ00ehiPL0d1oaTDfa14nFb3zI&redir_esc=y#v=onepage&q=duration&f=false, 1991.

Li, Z., Singhee, A., Wang, H., Raman, A., Siegel, S., Heng, F.-L., Mueller, R., and Labut, G.: Spatio-temporal forecasting of weather-driven damage in a distribution system, in: 2015 IEEE Power & Energy Society General Meeting, pp. 1–5, IEEE, 2015.

Lindsay, D. and Cox, S.: Effective probability forecasting for time series data using standard machine learning techniques, in: International Conference on Pattern Recognition and Image Analysis, pp. 35–44, Springer, 2005.

Liu, Y., Zhong, Y., and Qin, Q.: Scene Classification Based on Multiscale Convolutional Neural Network, IEEE Transactions on Geoscience and Remote Sensing, 56, 7109 – 7121, https://doi.org/10.1109/TGRS.2018.2848473, https://ieeexplore.ieee.org/abstract/document/8421052, 2018.

Mäkisara, K., Katila, M., Peräsaari, J., and Tomppo, E.: The multi-source national forest inventory of Finland–methods and results 2013, 2016.

Masson-Delmotte, T., Zhai, P., Pörtner, H., Roberts, D., Skea, J., Shukla, P., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., et al.: IPCC, 2018: Summary for Policymakers. In: Global warming of 1.5 C. An IPCC Special Report on the impacts of global warming of 1.5 C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global, Geneva, Switzerland, 2018.

myrskyvaroitus.com: 2018 Syystrombeja, https://myrskyvaroitus.com/index.php/myrskytieto/myrskyhistoria/202-2018-syystrombeja, 2018.

Nateghi, R., Guikema, S., and Quiring, S. M.: Power Outage Estimation for Tropical Cyclones: Improved Accuracy with Simpler Models, Risk Analysis, 34, 1069–1078, https://doi.org/10.1111/risa.12131, 2014.

Niemelä, E.: KESKEYTYSTILASTO 2017 (i), Tech. Rep. 2018-06-14 11:51:52.916, Energiateollisuus Ry, Eteläranta 10, 00130 Helsinki, Finland, https://energia.fi/files/2785/Sahkon_keskeytystilasto_2017.pdf, 2018.

Nurmi, V., Pilli-Sihvola, K., Gregow, H., and Perrels, A.: Overadaptation to Climate Change? The Case of the 2013 Finnish Electricity Market Act, Economics of Disasters and Climate Change, 3, 161–190, https://doi.org/10.1007/s41885-018-0038-1, 2019.

Pellikka, P. and Järvenpää, E.: Forest stand characteristics and wind and snow induced forest damage in boreal forest, in: Proceedings of the International Conference on Wind Effects on Trees, held in September, pp. 16–18, Citeseer, 2003.

Peltola, H., Kellomäki, S., and Väisänen, H.: Model Computations of the Impact of Climatic Change on the Windthrow Risk of Trees, Climatic Change, 41, 17–36, https://doi.org/10.1023/A:1005399822319, 1999.

Pinto, J. G., Bellenbaum, N., Karremann, M. K., and Della-Marta, P. M.: Serial clustering of extratropical cyclones over the North Atlantic and Europe under recent and future climate conditions, Journal of Geophysical Research: Atmospheres, 118, 12,476–12,485, https://doi.org/10.1002/2013JD020564, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JD020564, 2013.

Platt, J. et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers, 10, 61–74, 1999.

Ramon, J., Lledó, L., Torralba, V., Soret, A., and Doblas-Reyes, F. J.: What global reanalysis best represents near-surface winds?, Natural Disasters and Adaptation to Climate Change, 145, 3236–3251, https://doi.org/10.1002/qj.3616, 2019.

Rasmussen, C. E.: Gaussian processes in machine learning, in: Summer School on Machine Learning, pp. 63–71, Springer, 2003.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.02881, 2017.

Rossi, P. J.: Object-Oriented Analysis and Nowcasting of Convective Storms in Finland, Ph.D. thesis, Aalto University, http://urn.fi/URN:ISBN:978-952-60-6441-3, 2015.

Rupp, M.: Machine learning for quantum mechanics in a nutshell, International Journal of Quantum Chemistry, 115, 1058–1073, 2015.

Schelhaas, M.-J.: Impacts of natural disturbances on the development of European forest resources: application of model approaches from tree and stand levels to large-scale scenarios, Dissertationes Forestales, https://doi.org/10.14214/df.56, 2008.

Schelhaas, M.-J., Nabuurs, G.-J., and Schuck, A.: Natural disturbances in the European forests in the 19th and 20th centuries, Global Change Biology, 9, 1620–1633, https://doi.org/10.1046/j.1365-2486.2003.00684.x, https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2486.2003.00684.x, 2003.

Seidl, R., Schelhaas, M., Rammer, W., and Verkerk, P.: Increasing forest disturbances in Europe and their impact on carbon storage, Nature Climate Change, 4, 806–810, https://doi.org/10.1038/NCLIMATE2318, 2014.

Shawe-Taylor, J., Cristianini, N., et al.: Kernel methods for pattern analysis, pp. 296–297, Cambridge university press, 2004.

Shield, S. et al.: Predictive Modeling of Thunderstorm-Related Power Outages, Master's thesis, The Ohio State University, 2018.

Singhee, A. and Wang, H.: Probabilistic forecasts of service outage counts from severe weather in a distribution grid, in: 2017 IEEE Power & Energy Society General Meeting, pp. 1–5, IEEE, 2017.

Suvanto, S., Henttonen, H. M., Nöjd, P., and Mäkinen, H.: Forest susceptibility to storm damage is affected by similar factors regardless of storm type: Comparison of thunder storms and autumn extra-tropical cyclones in Finland, Forest Ecology and Management, 381, 17 – 28, https://doi.org/https://doi.org/10.1016/j.foreco.2016.09.005, http://www.sciencedirect.com/science/article/pii/S0378112716305266, 2016.

Tervo, R., Karjalainen, J., and Jung, A.: Short-Term Prediction of Electricity Outages Caused by Convective Storms, IEEE Transactions on Geoscience and Remote Sensing, 57, 8618–8626, 2019.

Ukkonen, P. and Mäkelä, A.: Evaluation of machine learning classifiers for predicting deep convection, Journal of Advances in Modeling Earth Systems, 11, 1784–1802, 2019.

Ukkonen, P., Manzato, A., and Mäkelä, A.: Evaluation of thunderstorm predictors for Finland using reanalyses and neural networks, Journal of Applied Meteorology and Climatology, 56, 2335–2352, 2017.

Ulbrich, U., Pinto, J. G., Kupfer, H., Leckebusch, G., Spangehl, T., and Reyers, M.: Changing Northern Hemisphere storm tracks in an ensemble of IPCC climate change simulations, Journal of Climaste, 21, 1669–1679, 2008.

Ulbrich, U., Leckebusch, G. C., and Pinto, J. G.: Extra-tropical cyclones in the present and future climate: A review, in: Theoretical and Applied Climatology, vol. 96, pp. 117–131, Springer Wien, https://doi.org/10.1007/s00704-008-0083-8, 2009.

Valta, H., Lehtonen, I., Laurila, T., Venäläinen, A., Laapas, M., and Gregow, H.: Communicating the amount of windstorm induced forest damage by the maximum wind gust speed in Finland, Advances in Science and Research, 16, 31–37, https://doi.org/10.5194/asr-16-31-2019, 2019.

Virot, E., Ponomarenko, A., Dehandschoewercker, Quéré, D., and Clanet, C.: Critical wind speed at which trees break, Physical Review E, 93, https://doi.org/10.1103/PhysRevE.93.023001, 2016.

Wang, G., Xu, T., Tang, T., Yuan, T., and Wang, H.: A Bayesian network model for prediction of weather-related failures in railway turnout systems, Expert Systems with Applications, 69, 247–256, https://doi.org/10.1016/j.eswa.2016.10.011, 2017.

Weather and Safety Center of Finnish Meteorological Institute - Duty forecasters: Personal communication, 05/2020.

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, pp. 76–99, Academic press, 2011.

Williams, C. K. and Rasmussen, C. E.: Gaussian processes for machine learning, vol. 2, MIT press Cambridge, MA, 2006.

Yang, F., Wanik, D. W., Cerrai, D., Bhuiyan, M. A. E., and Anagnostou, E. N.: Quantifying uncertainty in machine learning-based power outage prediction model training: A tool for sustainable storm restoration, Sustainability, 12, 1525, 2020.

Yue, M., Toto, T., Jensen, M. P., Giangrande, S. E., and Lofaro, R.: A Bayesian approach-based outage prediction in electric utility systems using radar measurement data, IEEE Transactions on Smart Grid, 9, 6149–6159, https://doi.org/10.1109/TSG.2017.2704288, 2018.

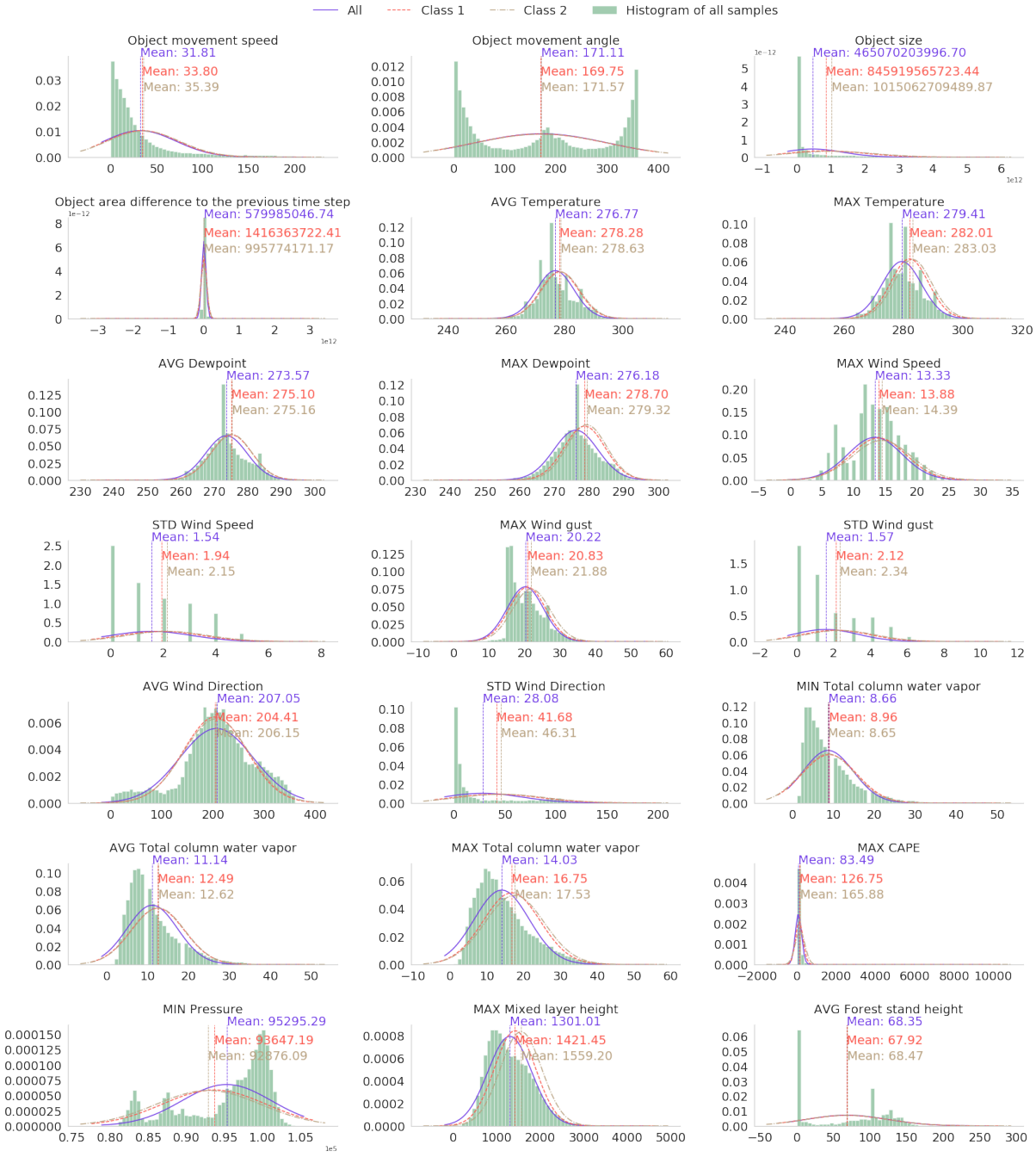


Figure A2. ~~Distribtution~~ Histogram of and fitted Gaussian distribution of selected predictive parameters ~~of~~ in **the national dataset**. The Gaussian distribution is fitted separately to all ~~storm objects (~~samples ~~)~~ and ~~objects~~ samples with ~~class 2. National dataset~~little outages and many outages (classes 1 and 2 specified in Section 3.3).