Response to the Comments RC2 on Manuscript Predicting power outages caused by extratropical storms

Corresponding author: Roope Tervo, roope.tervo@fmi.fi
Oct 7th, 2020

We are thankful for the precious and detailed comments and good improvement suggestions. We thank the reviewer for reading our paper carefully. In the following replies, we have addressed the comments as accurate and clear as possible and made the improvements in the manuscript.

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.

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

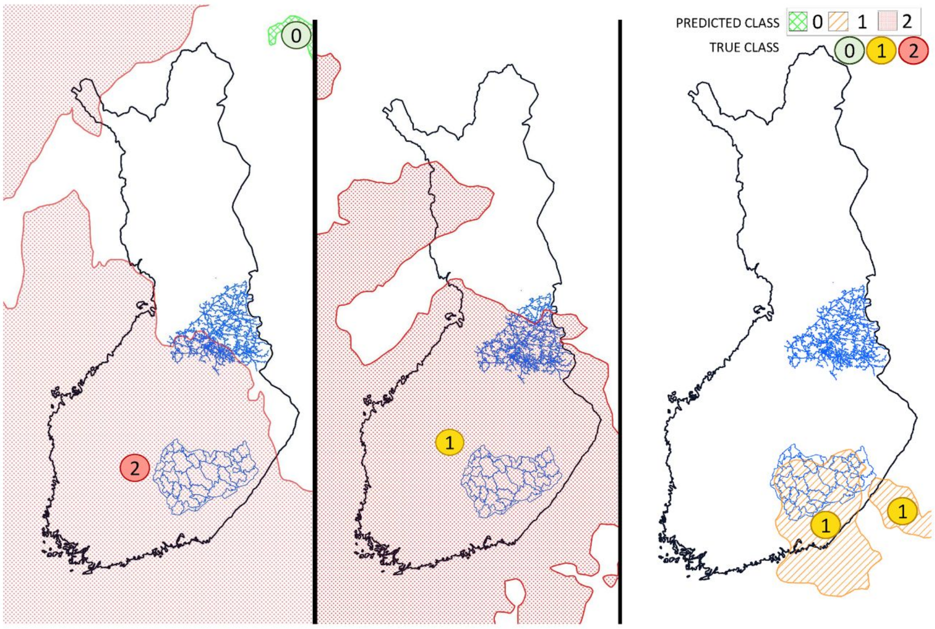
eastern or western part of the country and whether the damage in this region is expected to be light or severe. Moreover, in cases where a storm consists of several separate polygons, we are able to distinguish the damage potential of each polygon. Some examples of the coverages are illustrated in the 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 63-66:

“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. [...]”

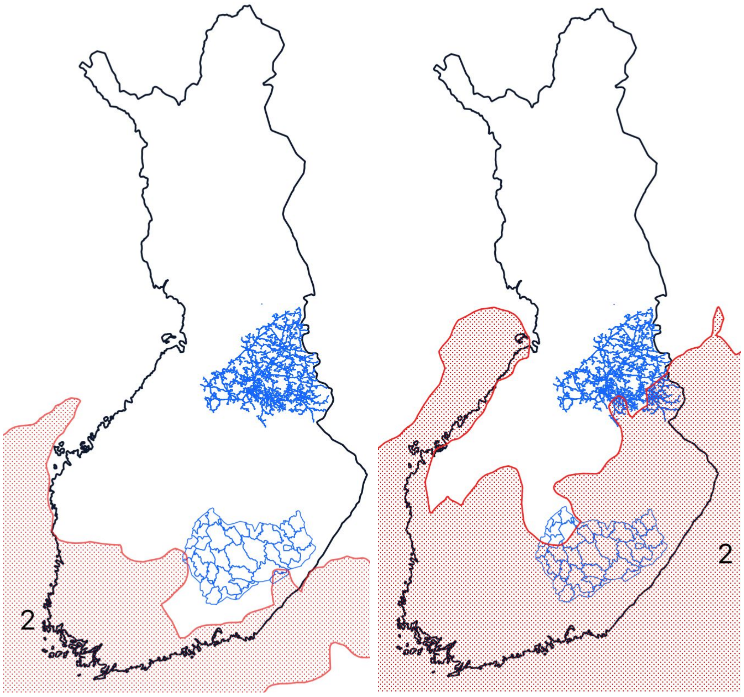
and in the chapter 3 Method on page 4, lines 117-122:

“We predict power outages by classifying storm objects identified from gridded weather data into three classes based on a number of power outages the storm can typically cause. The overall process contains the following steps: (1) identifying storm objects from weather fields by finding contour lines of some particular threshold, (2) tracking the storm object movement, (3) gathering features of the storm objects, and (4) classifying each storm object individually. The classification is conducted to each storm object separately to distinguish the different damage potential of each object. Tracking is, however, necessary to gather necessary features such as object movement speed and direction. Next, 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, the individual parameter, 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. Using completely independent features is not, however, possible in weather data since weather parameters are often dependent on each other, and eliminating even the most apparent pairs from used features impaired the results in our experiments.

We used the macro average of F1 defined in Equation 7 as a scoring function and 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 more significant 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) that are, however rather highlighting the importance of maximum wind gusts. 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 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 to the results much. 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).”

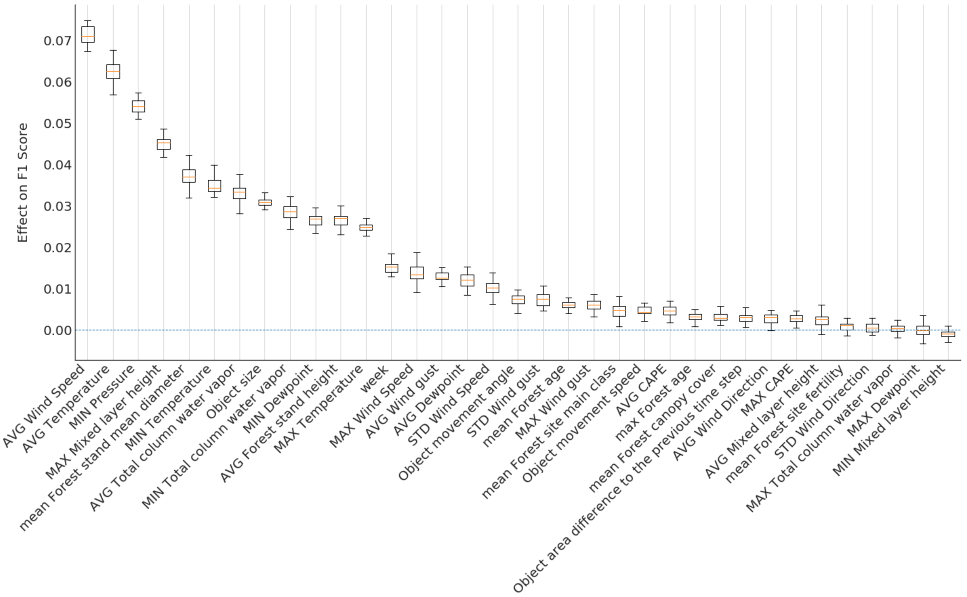


Figure7. Permutation feature importances using GP classification method trained with the randomly selected national dataset. The higher effect on the F1 score is (y-axis), the bigger is the significance.

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 290-293):

“[...] 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 23, lines 441-444):

"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 also in conclusion following (page 22, line 412-414):

“The evaluation was, however, based on the ERA5 reanalysis data. Using the method in operations 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 have increased and are increasing, while it remains unclear whether this is due to the increasing exposure of society or the number and intensity of extratropical storms. Gregow et al. (2017) discovered that windstorm damages have increased significantly during the past three decades, especially in northern, central, and western Europe. Also, several other studies suggest an increase in wind-related damages in Europe (Csillery 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 extratropical storm tracks to being sifted 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 Hersbach et al., 2019, 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 109-115. 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 from 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.”

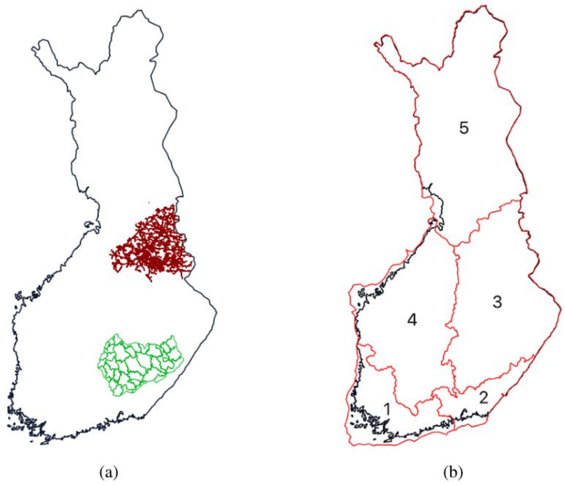


Figure 1. (a) Geographical coverage of the outage data (local dataset). The red lines represent the power grid of Loiste (northern grid company) and the green lines the operative areas of JSE (southern grid company). Outages of the local dataset are collected from both of these areas. (b) Regions in 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, 415-430:

“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 for wind gust speed to be the same for the almost whole 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 would 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 studies.”

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 135.

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 124-138:

“Storm objects are identified by finding contour lines of wind gust fields using 15 ms^{-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 hourly maximum wind gust exceeds 15 ms^{-1} during one particular hour. The threshold of 15 ms^{-1} is selected as different sources indicate Finland being vulnerable to windstorms and rather moderate winds (from 15 ms^{-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 ms^{-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 wind objects around the low-pressure center to the same storm event.

After identification, storm objects are connected to other storm objects around the common low-pressure objects and to the storm and pressure objects in preceding timesteps using Algorithm 1. Each object having pressure objects or preceding objects within the threshold is assigned to the same storm event and gets the same storm ID. Single storm objects without nearby pressure objects or preceding objects are left without ID as they are not assumed to be part of any storm.“

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 149-151.

“To represent each parameter with one number, we aggregate values as a minimum, maximum, average, and standard deviation calculated over all grid cells under the object coverage.”

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

Replaced with thanks on page 6, line 151.

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

Replaced with thanks on page 6, line 152.

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 on page 6, line 153.

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

Replaced with thanks on page 6, line 154.

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 202 (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 202 (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).

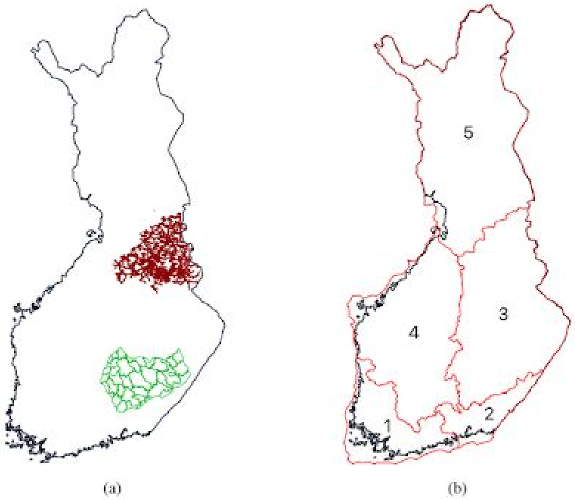


Figure 1. (a) Geographical coverage of the outage data (local dataset). The red lines represent the power grid of Loiste (northern grid company) district and the green lines the operative areas of JSE (southern grid company). Outages of the local dataset are collected from both of these areas. (b) Regions 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.

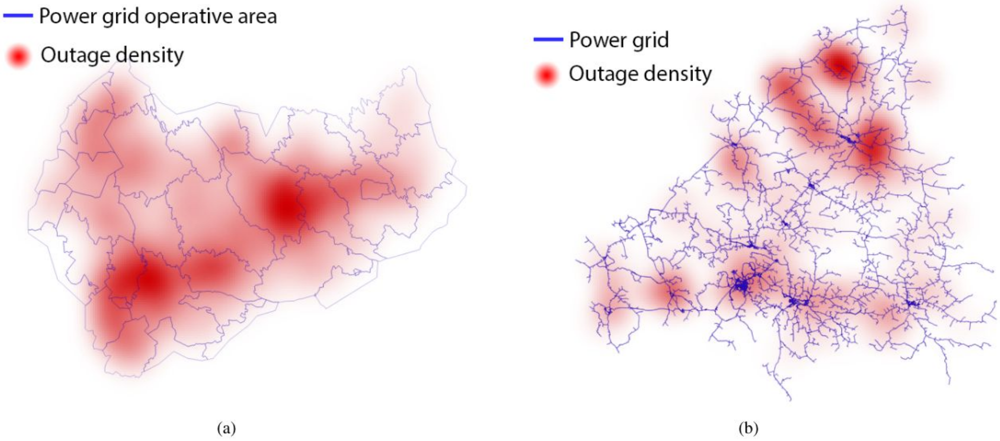


Figure 2. Spatial distribution of the outages between 2010 and 2018 visualised as a spatial heatmap. (a) JSE network (southern area) (b) Loiste network (northern area)

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 191-192.

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 12, lines 209-216:

“[...] 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 always to 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.

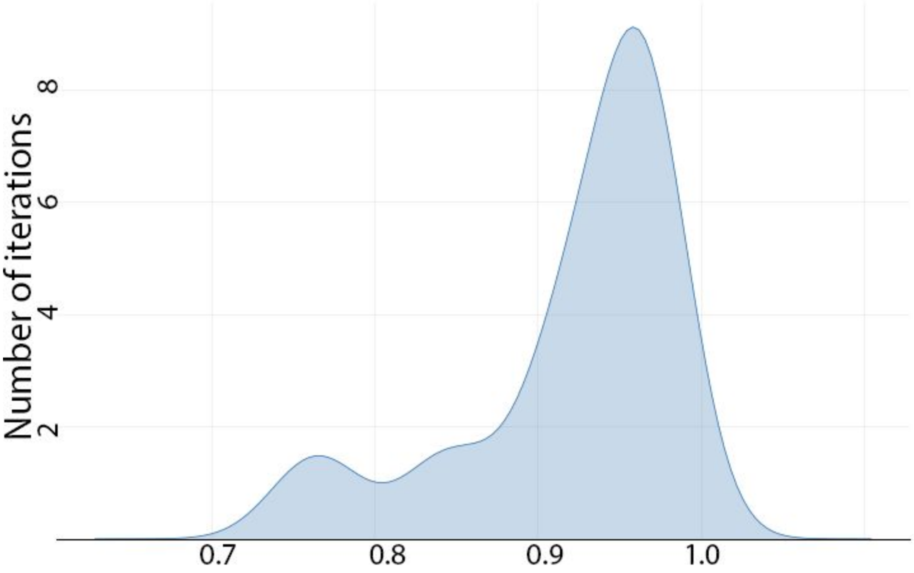


Figure: KDE plot of the results from the final iteration of hyperparameter and topology search.

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 19:

"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 412-414:
>
> “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 432-434:
>
> “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:

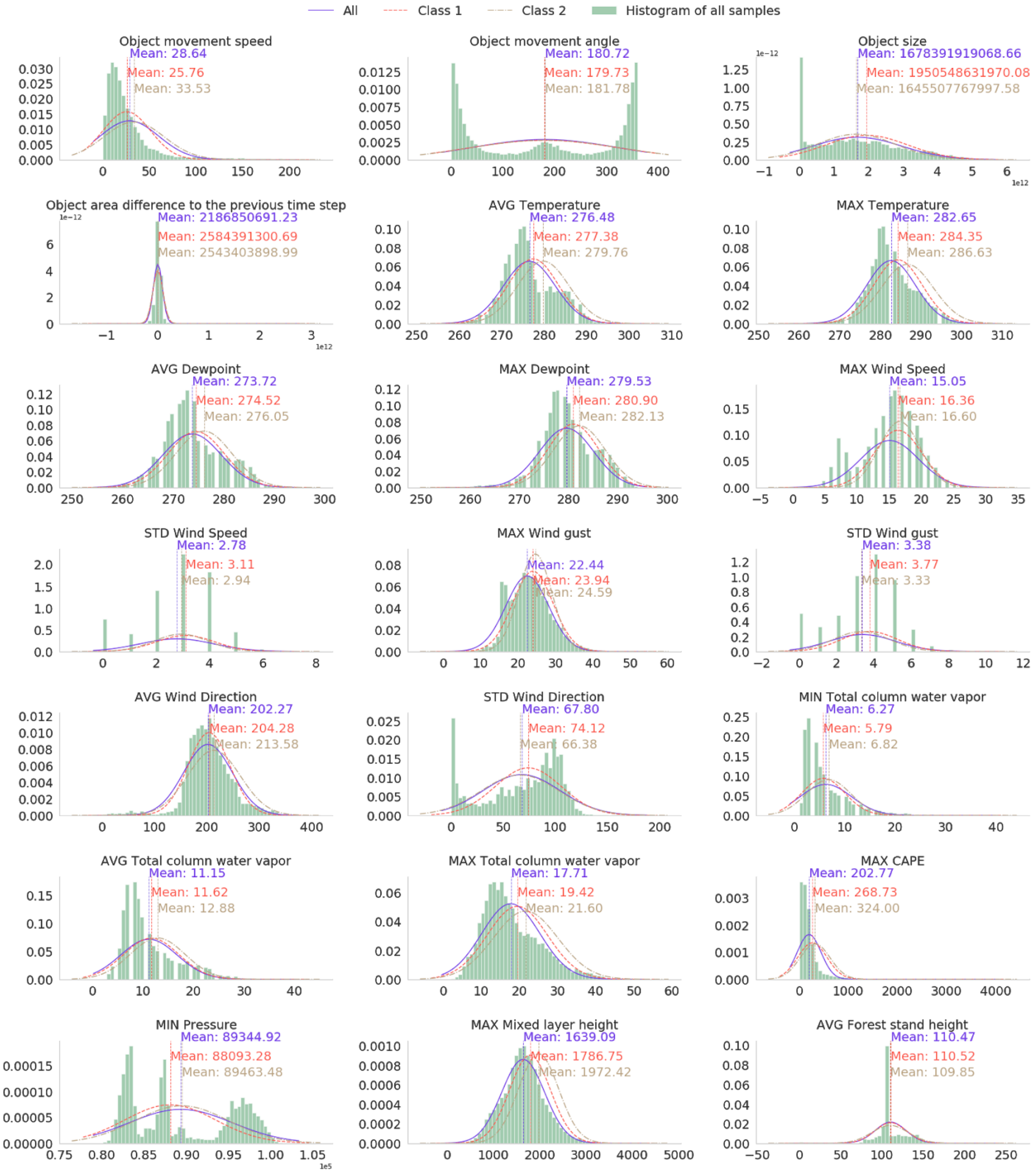


Figure A1. Histogram of and fitted Gaussian distribution of selected predictive parameters in the local dataset. The Gaussian distribution is fitted separately to all samples and samples with little outages and many outages (classes 1 and 2 specified in Section 3.3).

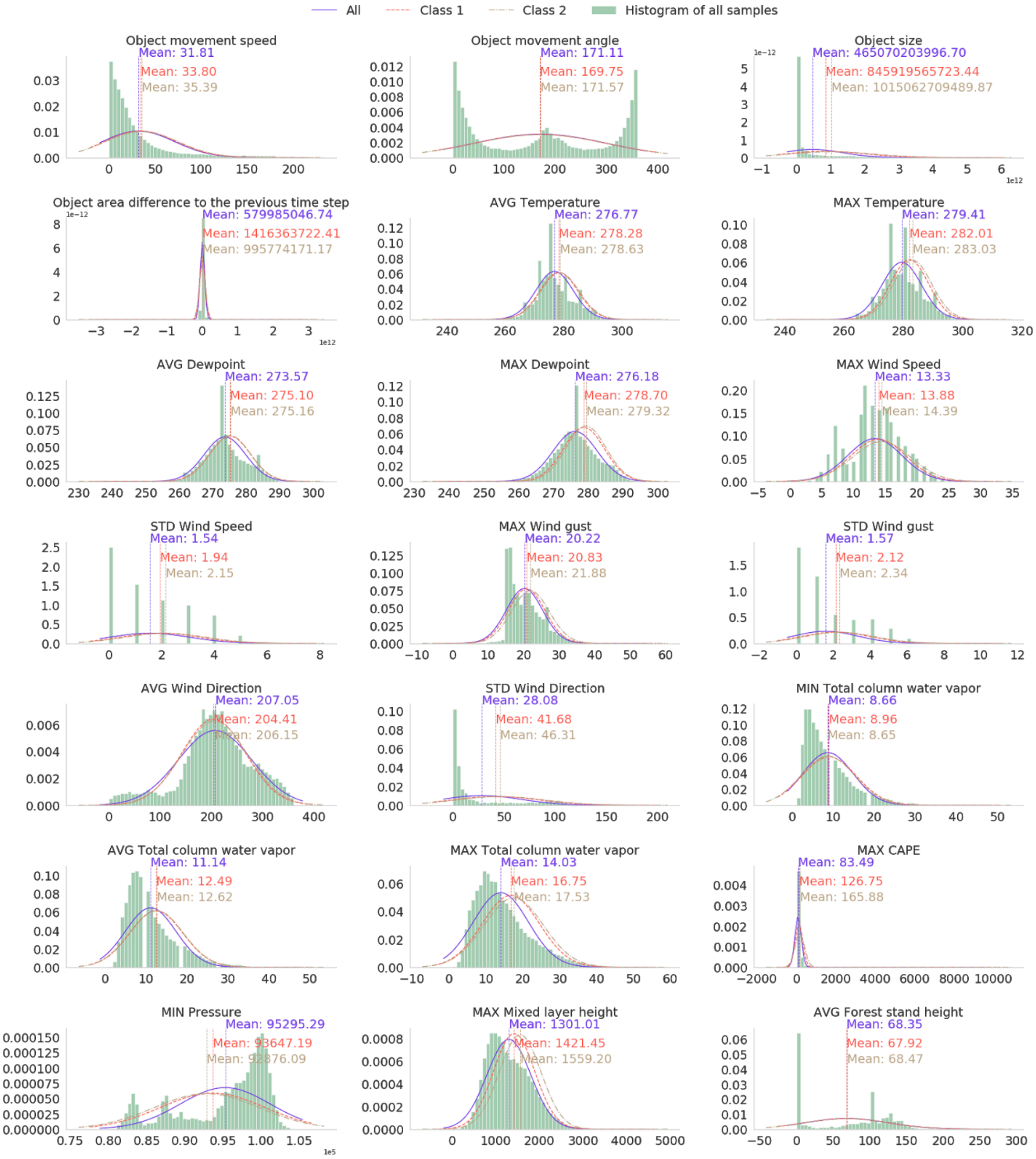


Figure A2. Histogram of and fitted Gaussian distribution of selected predictive parameters in the national dataset. The Gaussian distribution is fitted separately to all samples and samples with little outages and many outages (classes 1 and 2 specified in Section 3.3).