Response to the Comments SC1 on Manuscript Predicting power outages caused by extratropical storms

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

We are thankful and flattered for the commentator for reading our paper carefully and providing us valuable improvement suggestions. In the following replies, we have addressed the comments and made the improvements in the manuscript.

The commentators input is 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

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

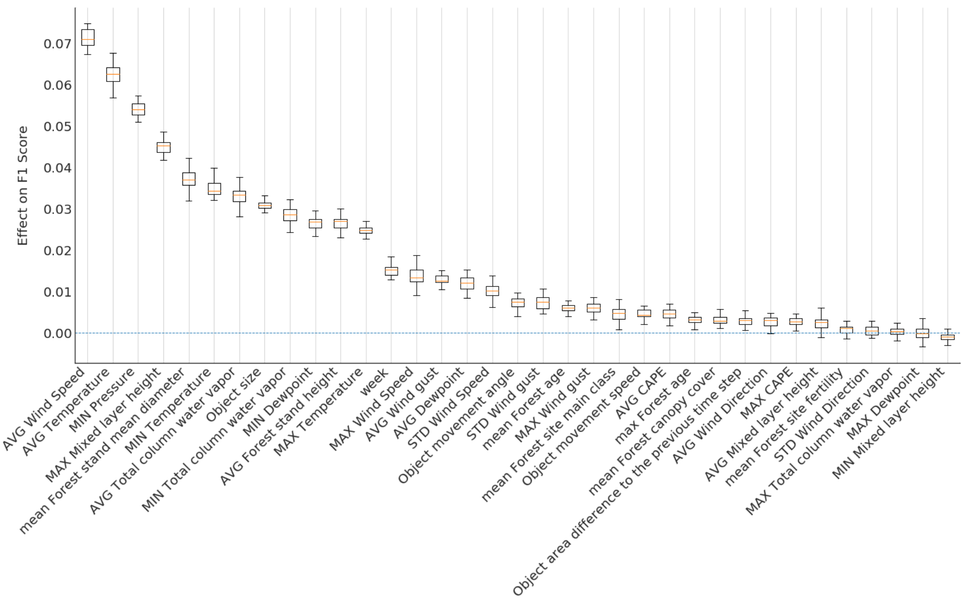
> We appreciate this advice and added mentioned papers to the previous work. (Updated manuscript page 2, lines 52-54).

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 198-199:
> "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.

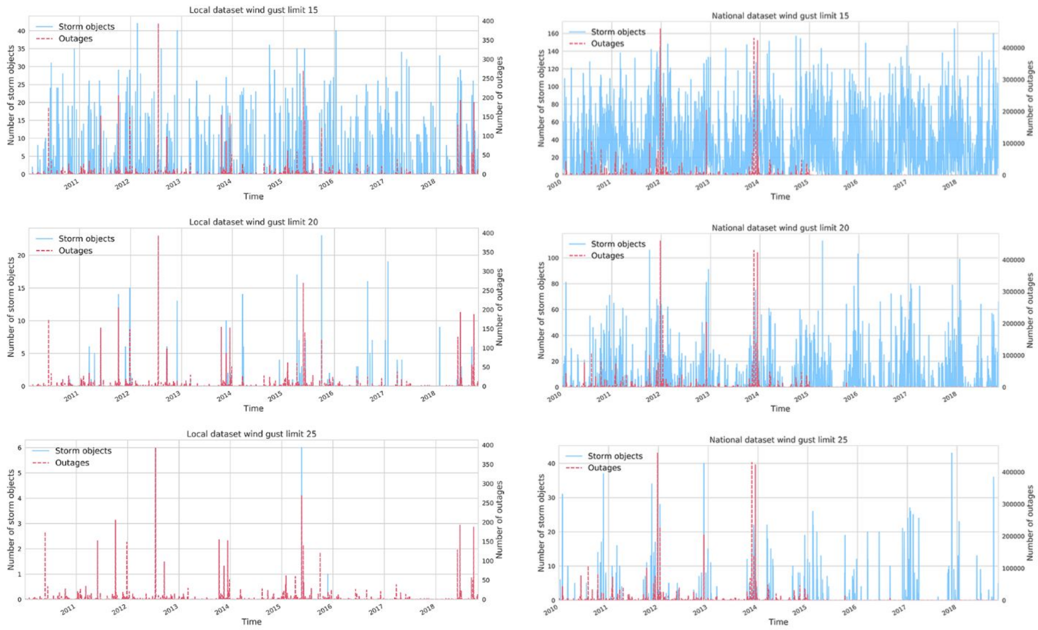


Figure 2. Storm object time series (15, 20 and 25 $\mathrm{m\ s^{-1}}$ contours) with occurred outages for local and national datasets.

Nevertheless, we commented the issue in the Discussion section (page 22, lines 436-438 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 continuous holdout from 2010 to 2011 for the test set. The evaluation could also be extended by, for example, 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."