Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019

# Spatially Aware Ensemble-Based Learning to Predict Weather-Related Outages in Transmission

Tatjana Dokic[1], Martin Pavlovski[2], Djordje Gligorijevic[2], Mladen Kezunovic[1], Zoran Obradovic[2]

[1]Department of Electrical and Computer Engineering
Texas A&M University
College Station, TX, U.S.A.

[2]Computer and Information Sciences Department
Temple University
Philadelphia, PA, U.S.A.

## Abstract

*This paper describes the implementation of a prediction model for real-time assessment of weather related outages in the electric transmission system. The network data and historical outages are correlated with a variety of weather sources in order to construct the knowledge extraction platform for accurate outage probability prediction. An extension of the logistic regression prediction model that embeds the spatial configuration of the network was used for prediction. The results show that the developed model manifests high accuracy and is able to differentiate an outage area from the rest of the network in 1 to 3 hours before the outage. The prediction model is integrated inside a weather testbed for real-time mapping of network outage probabilities based on incoming weather forecast.*

## 1. Introduction

Weather conditions present a major threat to electricity networks as 75% of power outages are either (1) directly caused by weather-inflicted faults (e.g., lightning, wind impact causing surrounding vegetation to contact transmission lines), or (2) indirectly by failures of equipment, caused partially by weather exposure (e.g. prolonged overheating or exposure to lightning-induced over-voltages) [1].

Due to recent weather trends, the number and frequency of power outages has dramatically increased [2]. This growth of grid outages and associated reliability deterioration is primarily due to severe weather caused by high wind, lightning, snow/storm, floods, etc., which is often driven by increased variability and extremes in seasonal weather patterns. The atmospheric conditions most conducive to severe weather are expected to increase [3-5], triggering increases in outage frequency and finally resulting in huge economic, social, and environmental risks to power systems and its customers.

Variety of studies have addressed the impact of extreme [6-8] and catastrophic [9,10] weather on power system infrastructure. The impacts of large-scale storms and hurricanes have been evaluated [6], while risk analysis has been performed for evaluation of wind storm impacts [7]. The work in [11] provides a statistical analysis of the spread of outages over an electric transmission network during severe weather events. The time-varying weight factors were introduced as a measure of weather impact on component failure rates and restoration times [12]. Historical weather data were correlated with historical outage data in order to develop a damage forecast model for restoration in [13].

Recently, the focus was on trying to improve the outage area prediction. The solution developed by the Weather Company [14] calculates the probability of an outage area based on the unfolding weather conditions. The UConn Outage Prediction Model [15] provides prediction for up to 3 days with 6-hour resolution. However, there are still many challenges in combining weather forecast with utility outage prediction as pointed out in [16]. The mentioned solutions are accurate in detecting outage areas, or predicting the number of expected outages in an area, but they are rather imprecise in identifying the exact outage locations. Obtaining a solution that is not only more accurate but also more stable remains a major challenge. To address the inaccuracy issue, the logistic regression model was used to predict weather related outage probabilities [17]. The solution in [17] was a good step to demonstrate the potential of using logistic regression to improve outage probability prediction, but it did not take advantage of the integration of real-time weather forecast or spatial information to improve the knowledge source.

The proposed method utilizes the knowledge from historical outage and weather data to provide accurate predictions 1-3 hours ahead. However, since spatial proximity plays an important role when it comes to outage occurrence prediction, the data holds a certain spatial structure that needs to be taken into account. Recently, ensembles that learn from structured data

URI: https://hdl.handle.net/10125/59784
ISBN: 978-0-9981331-2-6

have shown to be quite effective [18, 19]. Moreover, collaborative ensembles [20, 19] were proposed to enhance the performance of ensemble models by allowing their constituent components to interact. Therefore, the proposed method relies on a collaborative structured ensemble scheme [19] and extends its capabilities by: a) Adapting the objective function proposed in [19] to handle binary classification problems such as outage occurrence prediction. This objective strives to meet a proper balance between *underfitting* and *overfittting*, which is a fundamental challenge in machine learning; b) Employing multiple "local" Logistic Regression models (ensemble components) to learn different substructures and exchange information across their substructures in a manner that minimizes the objective function; and c) Providing probability estimates for outage occurrences in addition to the outage occurrence predictions.

This novel solution not only improves the accuracy when predicting outage occurrences, but also provides high accuracy in separating the areas in which outages did not occur.

## 2. Weather Testbed Architecture

To properly capture the continuously evolving weather impact on power systems, insights into the geographical layout of an electricity grid, as well as the evolving weather conditions need to be presented in a granular spatiotemporal framework. Moreover, spatially and temporally correlated measurements, coming from both utility infrastructure and weather data sources, need to scale to the temporal dynamics of the knowledge extraction process [17].

The Weather Testbed that supports integration of Big Data sources related to weather impacts on electric transmission and distribution is depicted in Fig. 1. The platforms in Fig. 1 are loaded with electric utility data and environmental data from a variety of data sources. The testbed is aiming at emulating the utility control center capabilities by providing the following components: 1) Storage and manipulation of the Big Data using iRODS [21]; 2) Spatial integration of heterogeneous data using ArcGIS [22]; 3) Temporal integration of real-time measurements using OSISoft PI [23]; 4) Supercomputing capabilities for execution of data processing, prediction algorithms, and optimization solutions; and 5) Visualization of real-time progression of weather threats and their impact on the network using an integrated ArcGIS and OSISoft PI platform, and Activu display [24].

For managing the big data access within the testbed, the Integrated Rule-Oriented Data System (iRODS) data management software is used. This system enables the following capabilities: 1) setting up iRODS zones for hosting the data, 2) project-wide data management for policy enforcement, 3) logging activities for later auditing, 4) sharing local and remote data for ease of access from a single user interface, and 5) data exchange between iRODS and public software repositories for optimization of resources.

ESRI ArcGIS is used for the spatial correlation of data. The data preprocessing and extraction of parameters for the prediction model is done using existing and custom-made ArcGIS tools and scripts. The visualization of results is done using both ArcGIS (spatial representation of results) and OSISoft PI (temporal representation of real-time results) visualization capabilities. The extensions to ArcGIS developed for our purposes allow integration and spatiotemporal correlation of standard data types and models describing power systems in addition to novel data sources such as weather data.

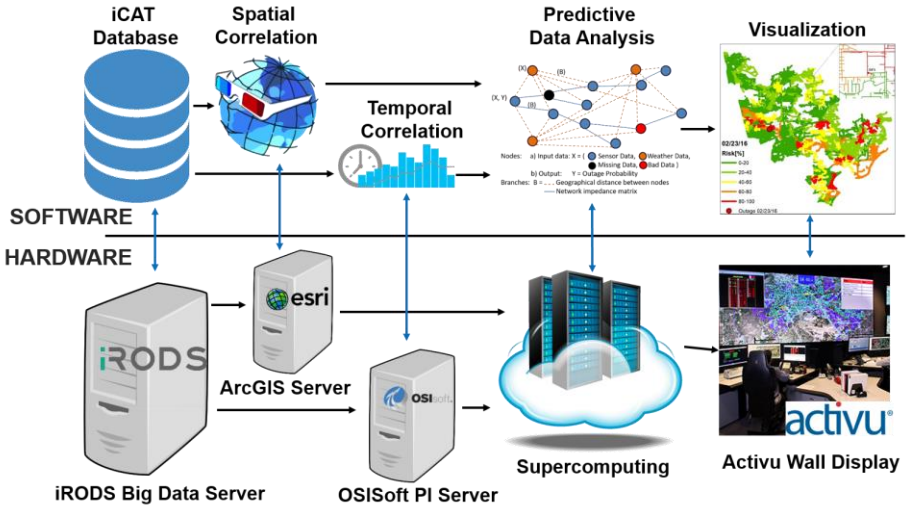


**Figure 1. Weather testbed architecture.**

The PI Historian platform is used for temporal analysis and visualization. Some of the data such as the weather station data (coming with resolution up to 1 min) and utility measurements are collected in real-time. This data is integrated using the OSISoft PI system. The goal of this study is prediction of weather related outages. Thus, the focus of the temporal data processing is to extract the parameters during the historical outages from the data collected in real-time.

The Activu wall display is used for visualization of the prediction model's results in real-time, emulating the real utility control center environment. The weather testbed allows for visualization of real-time weather risk maps that can enable transmission and distribution operators to follow the consequences of the unfolding environmental events on the severity of the impacts on the network.

This testbed demonstrates how the traditional sources describing different attributes of the power grid can be spatiotemporally associated with external sources of data and with the GIS and GPS features for improving decision-making capabilities. Such an architecture is capable of supporting a variety of weather related studies relevant to power system operation and planning, as well as to outage and asset management.

What differentiates this decision-making environment from the conventional utility solutions is the real-time processing and extraction of knowledge from unfolding weather forecasts for real-time interpretation of the impacts.

## 3. Data Sources and Processing

A variety of data sources was used for this study: 1) data from the utility geographical information system (GIS), 2) utility historical outage records, 3) historical weather measurements, 4) historical weather forecast data, and 5) elevation data.

Elevation data was extracted for the locations of all transmission substations using Elevation API provided by the Google Maps Platform [25]. The description of other data sources is provided in the following subsections.

**3.1. Historical Outages.** We used historical outage data from Bonneville Power Administration (BPA) [26]. The information for the transmission line outages caused by weather was extracted for the period from January 1st, 1999 to May 10th, 2018. A total of 16,806 weather related outages was identified. The following parameters were collected for each historical outage: 1) outage location, 2) outage time and date, 3) operating voltage, and 4) outage cause (lightning, ice, tree, tree cut, tree blown, tree growth, wind, earth slide, weather).

The geographical data for the BPA service area was obtained from [27]. As presented in Fig. 2, the following *shapefiles* were used: 1) *BPA_TransmissionLines*, 2) *BPA_Substations*, and 3) *BPA_ServiceArea*. A total of 639 substations were selected for the study. The network area spans over five states: Washington, Oregon, Idaho, Montana, and California.

**3.2. Weather Data.** For the extraction of weather parameters we used historical land-based weather station data collected by the Automated Surface Observing Systems (ASOS) program [28]. ASOS is a network of surface weather observations operating 24-hours a day with maximum temporal resolution of measurements of 1 min. The Iowa Environmental Mesonet (IEM) [29] was used for data download. A total of 84 weather stations were selected in the Pacific Northwest Area.

**Table I. Fractions of missing data from ASOS observations.**

| Temperature | DewPoint | RelHumidity | WindDirection | WindSpeed | Precipitation | Pressure | WindGust | WeatherCode |
|---|---|---|---|---|---|---|---|---|
| 0.146 | 0.148 | 0.148 | 0.145 | 0.134 | 0.312 | 0.265 | 0.378 | 0.336 |

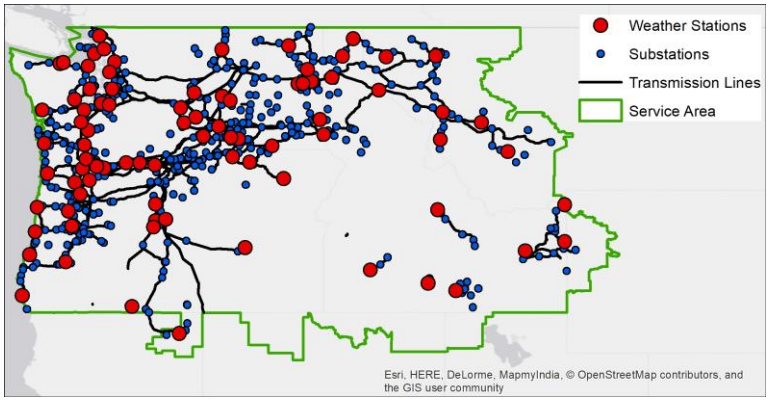


**Figure 2. Locations of ASOS weather stations and network components.**

The map of locations of weather stations across the network area is presented in Fig. 2. The following parameters were extracted from the ASOS data: Temperature [F], Dew Point [F], Relative Humidity [%], Wind Direction [degrees], Wind Speed [knots], Pressure [mb], Precipitation/Hour [inch], Wind Gust [knots], and Present Weather Codes. If there was no measurement of a parameter within 1 hour of the targeted time the value was declared missing. Table I lists the fractions of missing data, for each of the extracted parameters, out of a total of 34633 observation points (16806 with outages and 17827 without outages).

Weather forecast data was used for the construction of real-time outage probability maps. Historical weather forecast data was obtained from the National Digital Forecast Database (NDFD) [30]. The data was extracted using the NDFD GRIB Decoder - degrib [31]. The following elements were extracted: Temperature [F], Dew Point [F], Relative Humidity [%], Wind Direction [degrees], Wind Speed [knots], Precipitation Probability [%], and Wind Gust [knots]. The weather forecast for a time interval of 1-3 hours was extracted from the dataset for the time interval of 1-3 days from the Pacific Northwest NDFD Sector. An exception was the precipitation probability which is forecasted every 12 hours. The spatial resolution of forecast data is 5 km.

Before preprocessing, the total size of the historical weather dataset was ~14 GB (the weather stations in the vicinity of the network were selected for a period of 20 years). Weather forecast generates about 100 MB of data per day, which reaches about 35 GB for one year of testing.

## 4. Spatiotemporal Correlation of Data

The prediction model's input requires all the data sets to be spatiotemporally correlated. Fig. 3 presents the overview of this process. The first stage includes preprocessing and extraction of the ASOS, Outage, and Forecast tables individually. The second stage spatiotemporally correlates these tables into training, testing, and mapping datasets for further use by the prediction model.

The BPA geodatabase, containing locations of network substations, transmission lines and service area, was used as a spatial reference for the dataset extraction. The ASOS dataset was extracted from the IEM by selecting the required parameters for the weather stations in the network area, for the 1/1/1999-5/10/2018 period. The weather stations were selected based on their proximity to the network substations. The elevation data was extracted from the Google Maps Elevation API for the set of substation coordinates, and added to the outage table as an additional parameter.

The outage locations were extracted from the BPA outage table and correlated with the BPA map of transmission lines. The exact locations of outages are not known to the authors. The available transmission line historical outage dataset specifies the portion of transmission line where an outage occurred, but not the exact coordinates. For the purpose of easier processing and visualization of outage locations, every outage location was associated with its closest substation. This does not mean that the outage occurred in that substation, it means that the outage occurred in the close

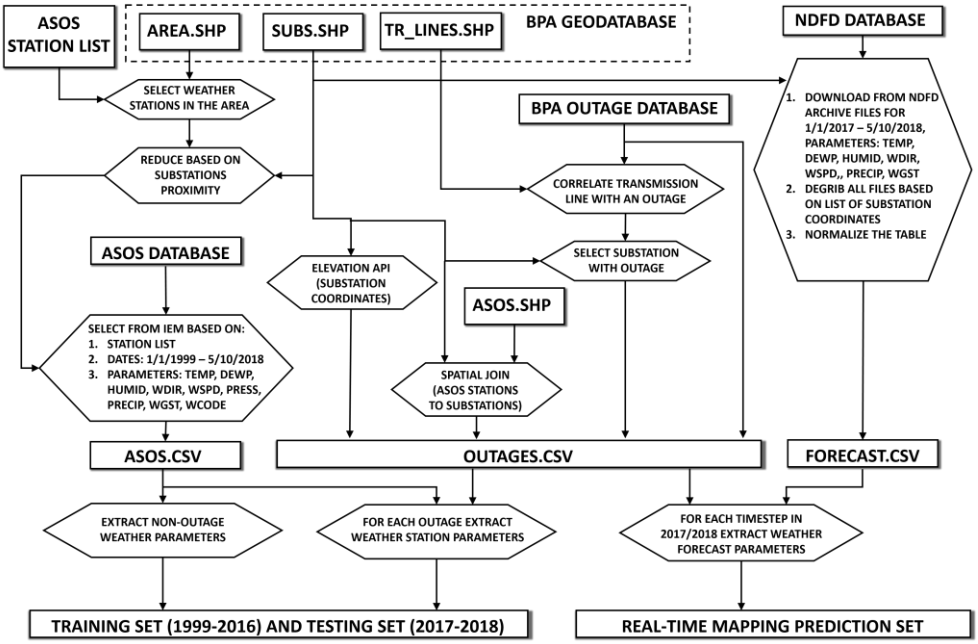


**Figure 3. Spatiotemporal correlation of data.**

vicinity to that substation (the selected substation is the closest substation to the outage).

For the spatiotemporal correlation of ASOS and outage data, the locations of ASOS stations were spatially joined to the substation locations and the associated ASOS station was added to every outage. Historical weather forecast data was used for the mapping of real-time outage prediction. For the number of time steps in 2017/2018 we downloaded the parameters of interest. Then the exact parameters for the times and locations of interest were extracted using the *degrid* function and the list of substation coordinates. The outcome of this first stage of processing are tree datasets, each containing detailed spatial and temporal reference: historical weather data from ASOS, 2) historical outage data from BPA, and 3) historical weather forecast from NDFD.

The second stage of processing creates training and testing datasets by extracting the ASOS parameters for each historical outage. In addition, a number of historical time steps without outages was extracted from the ASOS data so as to construct a balanced input dataset for the prediction model. The final training and testing datasets contain the following parameters: date and time, substation ID (location code), operating voltage, weather parameters from ASOS, presence of outage, and outage cause code. For the real-time mapping, the system weather forecast data was correlated with the historical outages. For multiple timesteps (some without outages, and some with different types of outages), we extracted the weather forecast made between 1 and 3 hours before the outage event based on the availability of NDFD data. For the probability of precipitation, the 12-hour forecast was extracted. The weather forecast does not contain pressure and weather codes. These parameters were removed from the prediction datasets for the purpose of real-time mapping.

The last part of the spatial analysis is visualization of results. For this purpose, the predicted outage probabilities for each substation were converted into a shapefile using the substation coordinates.

## 5. Prediction Model

The objective of this study is to estimate the probability of an outage event, given its location properties, time, operating voltage, and various weather-related parameters. Having an insight about the probability of such an event, action towards preventing an outage can be taken in a proactive manner. For this purpose, Logistic Regression [32], a probabilistic discriminative classifier, is considered for the task at hand. Formally, a Logistic Regression classifier models the posterior probability of an outage event occurrence $(y = 1)$, given a vector of measurements $\mathbf{x}$, as

$$P(y = 1|\mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}) ,$$

where $\mathbf{w}$ are the model's coefficients, and $\sigma$ is the *logistic sigmoid function*:

$$\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^{\mathrm{T}}\mathbf{x})} .$$

For the particular application of interest, let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T$ be a matrix in which each $\mathbf{x}_i \in \mathbb{R}^d$ (observation) contains features associated with the $i$-th substation. Moreover, let $\mathbf{y} = [y_1, \dots, y_N]^T$ be their corresponding class labels such that $y_i \in \{0,1\}$ is the label of $\mathbf{x}_i$. If $y_i = 0$ an outage event did not occur, whereas $y_i = 1$ indicates an outage occurrence, near the $i$-th substation.

A Logistic Regression model is fitted using the conditional distribution of the labels $\mathbf{y}$, given the observations $\mathbf{X}$:

$$P(\mathbf{y}|\mathbf{X}; \mathbf{w}) = \prod_{i=1}^{N} \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^{y_i}\big(1 - \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)\big)^{1-y_i} .$$

The model is fitted by determining the optimal coefficients $\mathbf{w}$ that maximize the logistic loss function, i.e.

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\, \mathcal{L}(\mathbf{w}) ;\ \mathcal{L}(\mathbf{w}) = \log P(\mathbf{y}|\mathbf{X}; \mathbf{w})$$
$$= \sum_{i=1}^{N} y_i \log \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i) + (1 - y_i) \log\big(1 - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)\big) .$$

The above optimization can be carried out by gradient-based methods [33] since $\mathcal{L}(\mathbf{w})$ is convex and its optimization is not constrained.

**5. 1. Accounting for spatial proximity by substation embedding.** The described logistic regression model is aimed at learning the relationship between the outage outcomes $y_i$ and the substations' features $\mathbf{x}_i$. However, a limitation of such a model is that it cannot account for the dependencies among the substations such as their spatial correlations described by the distances between them. For instance, if an outage occurs on a transmission line leaving a certain substation, it is likely that a nearby substation will record an outage as well. Such information is not captured by traditional probabilistic models such as the Logistic Regression model.

In our study, this limitation is addressed by learning representations of substations based on their spatial proximity. More precisely, the modularity approach [34] is used to generate vector representations (embeddings) in a $K$-dimensional space such that two substations that are spatially close to each other have similar representations.

Following a weighed graph-based formulation, nodes represent the substations while the links' weights are computed as distances between substations. Assuming that $G$ is a uniform random graph, the expected number of links between nodes $i$ and $j$ whose degrees are $d_i$ and $d_j$ is $\frac{d_i d_j}{2m}$, where $m$ is the total number of links in $G$. Therefore, the *modularity matrix* **B** for the adjacency matrix **A** is determined as

$$\mathbf{B} = \mathbf{A} - \frac{1}{2m}\mathbf{d}\mathbf{d}^T .$$

The matrix **B** is then decomposed using SVD and the top $K$ eigenvectors are used to embed the nodes (substations) in $G$.

These embeddings are appended to the original substations' features and the extended feature vectors are used to learn a Logistic Regression model as

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} \log P(\mathbf{y}|[\mathbf{X}, \mathbf{V}']; \mathbf{w}) ,$$

where the rows in $\mathbf{V}'$ are the substations' embeddings. This change in the input representation essentially adds an implicit spatial-awareness to the model's capability.

**5. 2. Collaborative Logistic Ensemble Classifier (CLEC).** A classification model built upon embeddings that incorporate the spatial proximity of substations, in addition to their original features, may capture the overall structure among them. However, such a model is not aware of useful substructures within the network of substations. An ensemble-based model was proposed in [19] to further capture hidden substructures within networks and, at the same time, aim at attaining the proper balance between *bias* and *variance*, and thus between *underfitting* and *overfitting*, by accounting for specific generalization insights in structured regression. Here, we extend the capabilities of this approach to the Collaborative Logistic Ensemble Classifier (CLEC) which can handle binary classification problems.

Consider the training dataset $\mathcal{D} = \{z_1 = (\mathbf{x}_1, y_1), \dots, z_N = (\mathbf{x}_N, y_N)\}$ in which the constituents of **X** and **y** are organized into pairs. The bias-variance balancing objective function of CLEC is defined as

$$R_{obj}(h, \mathcal{D}) = \sqrt{R_{emp}(h, \mathcal{D})^2 + dCorr(\ell(\cdot, h), z_{trn})^2} ,$$

where $R_{emp}(h, \mathcal{D}) = 1/N \sum_{i=1}^{N} \ell(z_i, h)$ is the empirical risk of a model $h$ w.r.t. $\mathcal{D}$ and $dCorr(\ell(\cdot, h), z_{trn})$ is the distance correlation [35], a measure of statistical dependence between a value outputted by a given loss function $\ell(\cdot, h)$ and a random training example (observation) $z_{trn}$. Essentially, minimizing the first term in $R_{obj}(h, \mathcal{D})$ protects against underfitting, while minimizing the second term indirectly prevents from overfitting ([19] explains this in more detail). Although $R_{obj}(h, \mathcal{D})$ has been initially proposed for structured regression problems, it can be easily generalized to a different supervised learning problem by defining $\ell(\cdot, h)$ to suit the problem at hand. As this study concerns a binary classification problem, the loss function is chosen to assess misclassifications, i.e. $\ell(z_i, h) = I\big(y_i \neq h(\mathbf{x}_i)\big)$, where $I$ is an indicator function.

To discover hidden data substructures, CLEC employs multiple "local" Logistic Regression models. For this purpose, $\mathcal{D}$ is sampled uniformly $M$ times using stratified sampling without replacement, thus generating $M$ data subsets $\mathcal{D}^1, \dots, \mathcal{D}^M$ of size $\eta N$, where $\eta \in (0,1)$. Thereafter, each $\mathcal{D}^m$ is used to train a single Logistic Regression component $F_{D^m}$. Upon training all $M$ components, the label of an unobserved substation $\mathbf{x}_s = [\mathbf{x}, \mathbf{v}']$ can be predicted as

$$\Phi_{\mathcal{D}}(\mathbf{x}_s) = \operatorname{sign}\left(\sum_{m=1}^{M} F_{D^m}(\mathbf{x}_s)\right).$$

As for the probability scores of $\Phi_{\mathcal{D}}$, they are taken to be the average (median can also be used) of the probabilities estimated by the components $F_{D^m}$.

Further, the components' subsets are modified by allowing the components to exchange information across their subsets. Essentially, the observations (each corresponding to a single substation) are exchanged among the components in a way that maximizes the difference between the values of $R_{obj}$, calculated before and after each exchange, i.e.

$$\begin{aligned}(j^*, k^*) &= \underset{(i,j)}{\operatorname{argmax}}\ \Delta_{jk} \\ &= \underset{(i,j)}{\operatorname{argmax}}\ R_{obj}(\Phi_{\mathcal{D}}, \mathcal{D}) - R_{obj}\big(\Phi_{\mathcal{D}^{(j,k)}}, \mathcal{D}\big) .\end{aligned}$$

## 6. Evaluation and Results

**6.1. Data Preprocessing**. The original data contained missing values in several, mostly weather-related, features (Table I provides the exact fractions of missingness per feature). To cope with this challenge, a nearest-neighbor imputation technique was used. Moreover, several features were constructed in addition to the original ones so as to better capture temporal dependencies among the substations. These include: days between ad-hoc measurements at substations; hour of day when measurements were performed, along with the season that day falls in. The hour of measurements was categorized within [0,23], while a one-hot representation was used to binarize the season feature. In addition, the elevation of each substation was pulled out and added as a separate feature. The rest of the

features (weather-related) were normalized using a min-max normalization, thus scaling them between 0 and 1.

**6. 2. Experimental Setup.** The experiments were conducted in a rigorous manner in terms of the horizon set for prediction. Namely, all models were trained using the data from 1999 to 2010, while future data up to 2018, totaling a prediction horizon of 9 years, was used for testing.

In all of the following experiments, the information from the spatial distance graph of substations has been embedded into a 50-dimenstional space using the modularity-based approach described in Section 5.1.

**6.2.1. Baseline Methods:** The prediction performance of CLEC was evaluated and compared against the following alternatives:

- **Logistic Regression (LR):** The use of this model for estimating the probability of outage occurrences was initially suggested in [17]. Moreover, since the proposed model incorporates multiple LRs as its components, LR was considered as a primary baseline.
- **Logistic Regression with *spatial* information:** This alternative utilizes the substations' spatial information by extending the original substations' feature vectors with spatial embeddings learned from the substation distance graph. Spatial information has also been shown to be beneficial when applied to similar tasks [36].

**6.2.2. Evaluation Metrics:** To assess the classification performance of predicting outage occurrences and the quality of their corresponding probability scores, the following metrics were considered: (1) *Accuracy* (ratio of correctly classified outages), (2) *Area under the Receiver Operating Characteristic curve (AUC)*, (3) *F1 score* (harmonic average of a model's accuracy w.r.t. both prediction of an outage occurrence when it did not occur, and vice-versa) and (4) *Bias* (expected misclassification error). The greater the value of Accuracy, AUC and F1, the better, while Bias gets better as it approaches 1.

**6.3. Outage Occurrence Prediction.** The prediction capability of CLEC was assessed for the task of classifying whether an outage event occurred at a certain substation or not based on the probability estimates of its occurrence. Its performance was assessed using the aforedescribed metrics. The obtained results are summarized in Table II.

First, from Table II, LR (spatial) obtains greater classification performance compared to LR which is consistent across all measures. This supports the hypothesis that spatial information is truly relevant to this task. Moreover, it can be observed that CLEC outperforms its alternatives, yielding higher values for Accuracy, AUC and F1. The large lift in Bias shows the benefit of using an ensemble-based model whose components focus on multiple data subsamples. The Bias of CLEC can be interpreted as having 232 expected misclassifications, out of a 1000, while LR (spatial), which is the next-best performing model, is expected to have 293. Thus, in theory, CLEC is expected to avoid ~61 outage occurrence misclassifications on every 1000 predictions.

**6.4. Logistic Model Coefficient Analysis.** To inspect the impact that spatial information has on prediction, the coefficient weights of LR and LR (spatial) were compared and presented in Fig. 4. It can be seen from the left hand side of the figure that the most relevant features to LR, w.r.t. their coefficients' magnitudes, are the last 3 features that correspond to precipitation, air pressure and wind gust. This is not a surprise, since these features are related to occurrence of severe storms that are one of the most dominant factors affecting the power outages. Once the spatial features are added (right hand side of Fig. 4), one can observe that their coefficients vary similarly to the coefficients of the original features. Finally, it was observed that spatial features contribute to 3 out of the top 10 largest coefficients of LR (spatial), thus showing that spatial information is significantly relevant.

**6.5. Performance Variability across Seasons.** The prediction performance of all models was also evaluated across different seasons (see Table III). The obtained results in terms of Accuracy indicate that CLEC consistently outperforms LR and LR (spatial), demonstrating improvements ranging from ~0.25-9.5% and ~0.33-6.2%, respectively. Improvements of CLEC in AUC and F1 are manifested in 3 out of 4 seasons. As for Bias, CLEC manifests improvements across all seasons. When compared to LR and LR (spatial), the expected outage occurrence misclassifications that can be avoided by CLEC range from 31-126 and 24-89 on every 1000 predictions, respectively. Overall, the largest improvements were achieved for the Winter season, while the smallest ones being recorded for the Summer season which reflects the volatility of the climate conditions of the region for which the data was collected and considered in this study.

**6.6. Real-time outage prediction mapping.** Fig. 5 shows the predicted real-time outage probability maps generated using weather forecast and a trained predictor. The maps were created for the timesteps presented in Table IV. The figures on the left show the results obtained using logistic regression, while the figures on

**Table II. Prediction performance w.r.t. different evaluation metrics.**

| Model | Acc. | AUC | F1 | Bias |
|---|---|---|---|---|
| LR | 0.8467 | 0.9278 | 0.8097 | 0.6821 |
| LR(spatial) | 0.8624 | 0.9292 | 0.8242 | 0.7075 |
| CLEC | **0.8919** | **0.9313** | **0.8532** | **0.7685** |

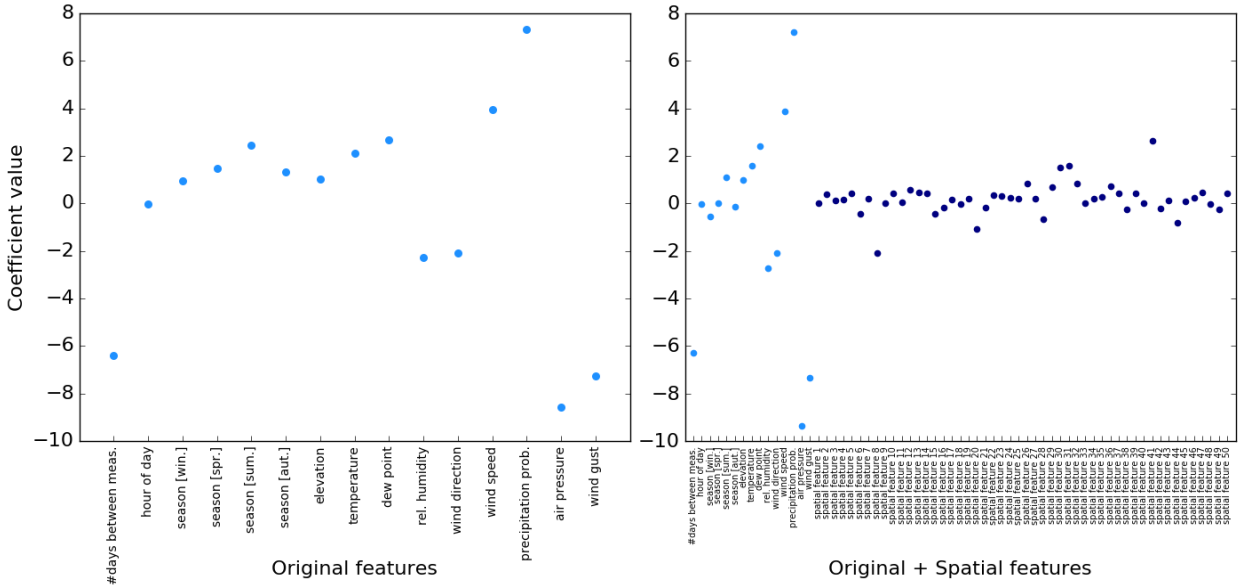


**Figure 4. Coefficients (or feature weights) assigned by a Logistic Regression model trained without (left) and with (right) spatial features; the magnitude of each coefficient value represents the importance of the feature it corresponds to w.r.t. predicting outage occurrences. The coefficient values corresponding to the original features are depicted using light blue, while the values corresponding to the spatial features are depicted using dark blue color.**

the right present the maps created using the proposed prediction model. The high risk locations in the network with over 80% outage probability (red color in Fig. 5) were enlarged for more convenient visualization.

The following observations can be made from the maps: 1) for the case when there was no outage in the network, the predicted outage probability was smaller than 60% for all substations; 2) for the cases when there was an outage, the area around the outages had points with very high probability (over 80%) and the rest of the network had no points with outage probability higher than 60%; 3) both logistic regression and the proposed prediction model are very good at guessing the area of the outage for all types of outages; 4) the proposed prediction model is better than logistic regression in terms of making prediction precession better on the spatial level (the number of high risk areas far away from the outage locations was much smaller).

Ideally, we would like to see red color at the location of outages, and dark green color everywhere else in the network. This is because we want to perform preventive actions only in the area of the outage, and not have to send maintenance crews all over the network. The proposed prediction model is closer to this goal than the logistic regression alone as can be seen by comparing the figures on the left and right in Fig. 5.

**Table III. Prediction performance w.r.t. different evaluation metrics across different seasons.**

| Model | Acc. | AUC | F1 | Bias |
|---|---|---|---|---|
| ***Winter*** | | | | |
| LR | 0.9089 | 0.8358 | 0.7340 | 0.5862 |
| LR spatial) | 0.9176 | 0.8451 | 0.7533 | 0.6272 |
| CLEC | **0.9305** | **0.8634** | **0.7803** | **0.7128** |
| ***Spring*** | | | | |
| LR | 0.8597 | 0.9361 | 0.8221 | 0.6687 |
| LR(spatial) | 0.8792 | 0.9325 | 0.8419 | 0.6932 |
| CLEC | **0.9164** | **0.9363** | **0.8822** | **0.7463** |
| ***Summer*** | | | | |
| LR | 0.7849 | 0.8860 | **0.8770** | 0.8540 |
| LR(spatial) | 0.7841 | 0.8843 | 0.8753 | 0.8613 |
| CLEC | **0.7874** | **0.8914** | 0.8766 | **0.8851** |
| ***Autumn*** | | | | |
| LR | 0.8132 | 0.8906 | 0.6855 | 0.5130 |
| LR(spatial) | 0.8462 | **0.8967** | 0.7211 | 0.5429 |
| CLEC | **0.9080** | 0.8874 | **0.7961** | **0.6312** |

## 7. Conclusion

Following are the contributions of this work:

- Logistic regression is extended with a generalization-aware structured learning of an ensemble in which the components interact by exchanging substations in a manner that strives to achieve a proper balance between underfitting and overfitting.
- The obtained solution is not only more accurate than the alternatives but is also more stable. It achieves improved accuracy of predicting outage

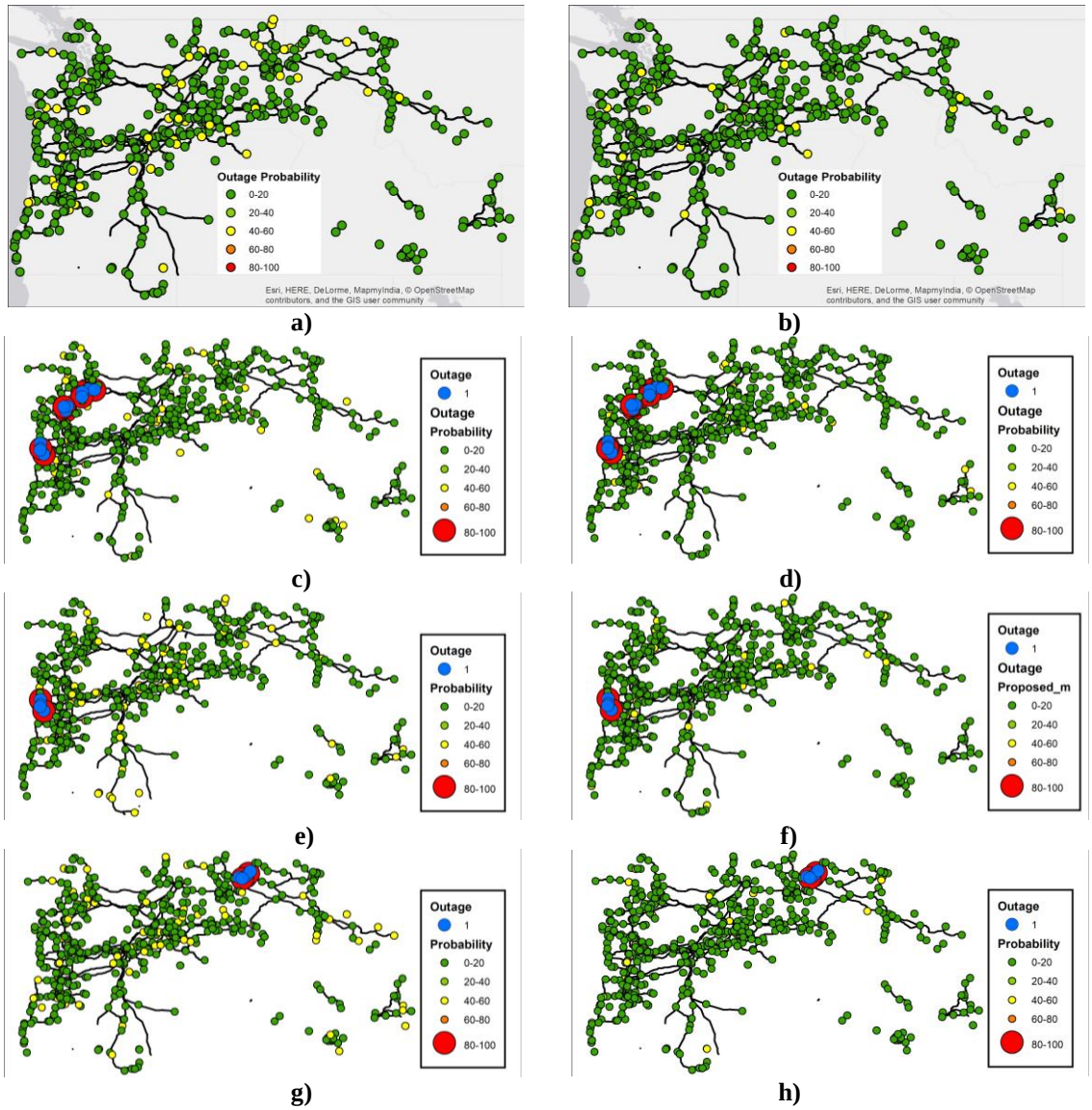


**Figure 5. Probabilities and locations of outages for: a) no outage - logistic regression, b) no outage - proposed method, c) lightning - logistic regression, d) lightning - proposed method, e) vegetation - logistic regression, f) vegetation - proposed method, g) ice - logistic regression, h) ice - proposed method.**

locations as well as of identifying areas without an outage.

- The spatial structure of the utility network is embedded into the logistic regression prediction model for improved spatial granularity of prediction and localization of outages.
- The proposed model shows high accuracy of identifying outage locations for the weather forecast of 1 to 3 hours in advance of an event.

**Table IV. Historical cases for the real-time mapping example.**

| Timestep | Start timestep time | End timestep time | Presence of outage | Outage cause | Figure |
|---|---|---|---|---|---|
| 5 | 5/1/2017 12:00 | 5/1/2017 15:00 | 0 | NA | a) b) |
| 32 | 5/4/2017 21:00 | 5/5/2017 0:00 | 1 | lightning | c) d) |
| 1133 | 9/19/2017 12:00 | 9/19/2017 15:00 | 1 | vegetation | e) f) |
| 1866 | 12/20/2017 3:00 | 12/20/2017 6:00 | 1 | ice | g) h) |

- The weather testbed environment for integrati on of weather datasets into the utility control center is developed and described. This kind of environment allows seamless integration of weather data into all applications of interest to utility operation.
- Methods for extraction and spatiotemporal correlation of variety of datasets are implemented, including BPA outage and GIS data, ASOS weather station data, and NDFD weather forecast data.
- A real-time mapping system is developed for observing outage probabilities in the network using weather forecast.

## 8. References


[1] Eaton, "Blackout Tracker United States Annual Report 2015," Eaton, 2016. [Online] Available: http://www.sustainablepowersystems.com/wp-content/uploads/2016/03/US_BlackoutTracker_2015_Final.pdf

[2] Executive Office of the President, "Economic Benefits of Increasing Electric Grid Resilience to Weather Outages," Aug. 2013. [Online] Available: http://energy.gov/sites/prod/files/2013/08/f2/Grid%20Resiliency%20Report_FINAL.pdf

[3] R. J. Trapp, et al., "Changes in severe thunderstorm environment frequency during the 21st century caused by anthropogenically enhanced global radiative forcing," PNAS, vol. 104, no. 50, pp. 19719-19723, doi:10.1073/pnas.0705494104, 2007.

[4] A. D. Del Genio, et al., "Will moist convection be stronger in a warmer climate?" Geophys. Res. Lett., Vol. 34, No. 16, doi:10.1029/2007GL030525, 2007.

[5] N. S. Diffenbaugh, et al., "Robust increases in severe thunderstorm environments in response to greenhouse forcing," Proceedings of the National Academy of Sciences, vol. 110, pp. 16361-16366, doi:10.1073/pnas.1307758110, 2013.

[6] D. Lubkeman, and D. E. Julian. "Large scale storm outage management." Power Engineering Society General Meeting, 2004. IEEE. IEEE, 2004.

[7] G. Li, et al. "Risk analysis for distribution systems in the northeast US under wind storms." IEEE Transactions on Power Systems, vol. 29, no. 2, pp. 889-898, 2014.

[8] D. Yates, et al. "Stormy weather: Assessing climate change hazards to electric power infrastructure: A Sandy case study." IEEE Power and Energy Magazine, vol. 12, no. 5, pp. 66-75, 2014.

[9] M. Panteli, et al. "Power System Resilience to Extreme Weather: Fragility Modelling, Probabilistic Impact Assessment, and Adaptation Measures." Accepted to IEEE Transactions on Power Systems, 2016.

[10] C. Chen, et al., "Modernizing Distribution System Restoration to Achieve Grid Resiliency Against Extreme Weather Events: An Integrated Solution," accepted to Proceedings of the IEEE, 2017.

[11] I. Dobson, et al., "Exploring Cascading Outages and Weather via Processing Historic Data," Proceedings of the 51st Hawaii International Conference on System Sciences, Waikoloa Village, Hawaii, January 2018.

[12] P. Wang, and Roy Billinton. "Reliability cost/worth assessment of distribution systems incorporating time-varying weather conditions and restoration resources." IEEE Transactions on Power Delivery, vol. 17, no. 1, pp. 260-265, 2002.

[13] L. Treinish et al., "Operational utilization and evaluation of a coupled weather and outage prediction service for electric utility operations," in Proc. 2nd Conf. Weather Climate New Energy Economy, Seattle, WA, USA, Jan. 2011.

[14] The Weather Company, "Outage Prediction," [Online] Available: https://business.weather.com/products/outage-prediction

[15] Eversource Energy Center, "Predicting Outages," [Online] Available: http://www.eversource.uconn.edu/storm-outage-forecasting/predicting-outages

[16] R. O. Mueller, J. Singer, "Electric Utility Outage Prediction Models: Assessing Their Accuracy & Implementing Improvements" Electric Energy Online, [Online] Available: http://www.electricenergyonline.com/energy/magazine/1037/article/Electric-Utility-Outage-Prediction-Models-Assessing-Their-Accuracy-Implementing-Improvements.htm

[17] M. Kezunovic, et al., "Systematic Framework for Integration of Weather Data into Prediction Models for the Electric Grid Outage and Asset Management Applications," Proceedings of the 51st Hawaii International Conference on System Sciences, Hawaii, January 2018.

[18] Pavlovski, Martin, et al. "Adaptive Skip-Train Structured Regression for Temporal Networks." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2017.

[19] Pavlovski, Martin, et al. "Generalization-Aware Structured Regression towards Balancing Bias and Variance." IJCAI. 2018.

[20] Arsov, Nino, et al. "Generating highly accurate prediction hypotheses through collaborative ensemble learning." Scientific reports 7 (2017): 44649.

[21] iRODS, "Open Source Data Management Software," [Online] Available: https://irods.org/

[22] ArcGIS, Esri. 2017 [Online] Available: https://www.arcgis.com

[23] OSISoft, 2017 [Online] Available: http://www.osisoft.com/

[24] Activu, "The Wall Display," [Online] Available: https://activu.com/products/activu-5/

[25] Google, Google Maps Platform, Elevation API, [Online] Available: https://developers.google.com/maps/documentation/elevation/start

[26] Bonneville Power Administration, "Miscellaneous Outage Data and Analysis," [Online] Available: https://transmission.bpa.gov/Business/Operations/Outages/default.aspx

[27] Bonneville Power Administration, "GIS data" [Online] Available: https://hub.arcgis.com/items?tags=Bonneville%20Power%20Administration

[28] NOAA, "Automated Surface Observing System (ASOS)" [Online] Available: https://www.ncdc.noaa .gov/data-access/land-based-station-data/land-based-datasets/automated-surface-observing-system-asos

[29] Iowa State university, "Iowa Environmental Mesonet," [Online] Available: https://mesonet.agron.iastate.edu/request/download.phtml

[30] National Weather Services, "NDFD," [Online] Available: https://www.weather.gov/mdl/ndfd_home

[31] National Weather Services, "Degrib Man Page," [Online] Available: https://www.weather.gov/mdl/degrib_manpage_degrib

[32] K. Murphy, "Logistic regression," Machine Learning: A Probabilistic Perspective, Chapter 8, pp. 245 – 279, 2012.

[33] S. Boyd, and L. Vandenberghe, "Convex Optimization," 2004, ISBN: 9780521833783.

[34] Newman, Mark EJ. "Modularity and community structure in networks." Proceedings of the national academy of sciences 103.23 (2006): 8577-8582.

[35] Székely, Gábor J., et al. Measuring and testing dependence by correlation of distances. The annals of statistics, 35(6):2769–2794, 2007.

[36] Dokic, et al., "Risk Assessment of a Transmission Line Insulation Breakdown due to Lightning and Severe Weather," Proc. HICCS – Hawaii International Conference on System Science, Kauai, Hawaii, January 2016.pp. 2488 - 2497.