Article

Quantifying Uncertainty in Machine Learning-Based Power Outage Prediction Model Training: A Tool for Sustainable Storm Restoration

Feifei Yang ¹, David W. Wanik ², Diego Cerrai ¹ , Md Abul Ehsan Bhuiyan ³ and Emmanouil N. Anagnostou ^{1,*}

¹ Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT 06269, USA; feifei.yang@uconn.edu (F.Y.); diego.cerrai@uconn.edu (D.C.)

² Department of Operations and Information Management, University of Connecticut, Stamford, CT 06901, USA; dave.wanik@uconn.edu

³ Department of Natural Resources and the Environment, University of Connecticut, Storrs, CT 06269, USA; md.bhuiyan@uconn.edu

* Correspondence: manos@uconn.edu

Received: 18 January 2020; Accepted: 14 February 2020; Published: 18 February 2020

Abstract: A growing number of electricity utilities use machine learning-based outage prediction models (OPMs) to predict the impact of storms on their networks for sustainable management. The accuracy of OPM predictions is sensitive to sample size and event severity representativeness in the training dataset, the extent of which has not yet been quantified. This study devised a randomized and out-of-sample validation experiment to quantify an OPM's prediction uncertainty to different training sample sizes and event severity representativeness. The study showed random error decreasing by more than 100% for sample sizes ranging from 10 to 80 extratropical events, and by 32% for sample sizes from 10 to 40 thunderstorms. This study quantified the minimum number of sample size for the OPM attaining an acceptable prediction performance. The results demonstrated that conditioning the training of the OPM to a subset of events representative of the predicted event's severity reduced the underestimation bias exhibited in high-impact events and the overestimation bias in low-impact ones. We used cross entropy (CE) to quantify the relatedness of weather variable distribution between the training dataset and the forecasted event.

Keywords: cross entropy; event severity representativeness; machine learning; outage prediction model; sample size; severe weather; uncertainty

1. Introduction

In the United States, power delivery interruptions caused by weather events cost the national economy billions of dollars annually and affect the lives of thousands of utility customers [1]. A best practice for reducing the amount of outages and their duration is for a utility to improve its storm resilience through storm preparedness and infrastructure reinforcement. These two resilience components can be studied using outage prediction models (OPMs) that represent the interaction among weather, environmental conditions, and electric infrastructure. In particular, the use of an OPM as a decision support tool for storm preparedness can help utility incidence control managers with their planning of crew and equipment allocation, resulting in faster power restoration.

Storm-based OPM, using machine learning (ML) methods and slicing an electric service territory into geographical grids, has been investigated since the early 2000s. Han et al. [2] studied five hurricanes in the central Gulf Coast region and used a generalized linear model (GLM) to predict power outages. However, the model tended to overestimate the number of outages in the urban areas

and underestimate those in rural regions. Han et al. [3] improved their OPM by using a Poisson generalized additive model (GAM), but it still over- and underestimated power outages in some areas. Subsequently, Nateghi et al. [4] used a random forest (RF) model to predict hurricane power outages based on the same dataset as Han et al. [2,3]. Their model tended to overestimate outages in some parts of the service territory. Guikema et al. [5] developed a hurricane outage prediction model by training 10 past hurricane events. This model overestimated outages in Maryland and Delaware and underestimated outages in Connecticut for hurricane Sandy. Wanik et al. [6] studied 89 severe storms from different seasons and applied four regression tree models—RF, decision tree (DT), boosted gradient tree (BT), and ensemble decision tree (ENS)—to predict power outages in the northeastern United States. He et al. [7] continued this work using quantile regression forests (QRF) and Bayesian additive regression tree (BART) models to predict power outages in the Northeastern United States. Cerrai et al. [8] used 76 extratropical and 44 convective storms, introduced the methods of storm type classification, and further developed the OPM based on the work of Wanik et al. [6] and He et al. [7]. In these last three studies [6–8], underestimation bias was exhibited in high-impact events and overestimation bias in low-impact events. Our model further advanced the models developed by Wanik et al. [6,8] and quantified the uncertainty of OPMs.

Since ML models learn patterns from a training dataset and apply them to a testing set, a good historical training record is considered essential to their predictive performance. To be successful, the training of ML models often requires a substantial amount of representative data [9]. The literature highlights many empirical examples of small training samples that have generally not performed as well as larger ones [10–12]. Although weather data can be available for many years of weather simulations, outage records may be more limited (covering two to five years). Additionally, given that only a handful of storms that cause widespread outages typically happen each year, a utility may need to rely on a limited number of events to train an OPM. Quantifying the uncertainty of OPMs associated with the training sample's number of historical events would help utilities know better the number of events they need to attain a required model performance. Studies in current ML-based outage modeling did not quantify the uncertainty of OPMs associated with the training sample sizes, since most have relied on limited sample sizes to demonstrate OPM improvements. For example, the hurricane studies mostly used only 10 storms in their models [2–5], while Wanik et al. [6] and He et al. [7] used less than 100 events in their OPM. Although Cerrai et al. [8] used 120 storms and selected the most important variables for their model, they did not quantify the uncertainty of OPMs with different sample sizes. In addition to the uncertainty associated with the training sample sizes, OPMs exhibit, as mentioned, overestimation biases of low-impact events and underestimation biases of high-impact events [6–8]. Given that the standard approach has been to use all the historical events as the training dataset, this may be the cause of prediction bias.

In summary, studies in current ML-based outage modeling have not adequately investigated the issue of uncertainty due to the training dataset sample size and the issue of low-impact event overestimation and high-impact event underestimation. These issues affect the reliability and use of the model predictions by utility managers. This study contributes to our understanding of these two issues by investigating the impact of event sample sizes and representativeness of predicted event severity in OPM training.

(1) Overcoming the uncertainty due to limited sample sizes: we used an unprecedented number (141) of events to quantify the association between prediction uncertainty and training event sample sizes. Understanding the training sample sizes effect on OPM uncertainty based on the Eversource-Connecticut data, could provide information in developing an OPM in other electric utility service territories which have limited data. This will help utilities to know the minimum number of training events needed to reach an acceptable model performance.

(2) Overcoming underestimation in severe events and overestimation in weak events: we explored training data classification to address prediction biases in high- and low-impact events caused by a lack of event severity representativeness. Specifically, we investigated bias reduction of weak and high

severity events by sub-setting the OPM's training dataset to historical events with outage severity similar to that of the predicted event. This approach is contrary to the conventional machine learning training methods, which use all available data.

The study was organized as follows. In the first part, we quantified the uncertainty of an OPM associated with different event sample sizes. We explored its accuracy by varying the number of storms used for training for extratropical and thunderstorm events, respectively, and conducted repeated randomized and out-of-sample validation experiments. Although our research focused on the Eversource-Connecticut service territory, understanding OPM uncertainty for each storm type associated with different training sample sizes could provide information to utilities developing an OPM in other service territories with similar infrastructure and vegetation characteristics. Furthermore, from this study we could determine the minimum number of the sample size for the OPM attaining an acceptable model performance for thunderstorm and extratropical events.

For the second part of the study, we focused on reducing the prediction bias by sub-setting the OPM training dataset according to events representative of the predicted event's severity. We defined event severity based on the number of outages in Connecticut, classifying the storms into three groups: low-impact events (below 200 outages), moderate-impact events (200 to 1000 outages), and high-impact events (above 1000 outages). These ranges correspond to thresholds a utility company might use for pre-storm allocation of crews. We selected 12 events as test cases from these three groups by stratified sampling [13,14]. The training dataset we used comprised all 92 extratropical events, and each test case was holdout in the training. Throughout this part of the study, we evaluated the reduction in the underestimation of high-impact and the overestimation of low-impact events from a training OPM to a more representative subsample of events.

2. Materials

The study area was the Connecticut service territory of Eversource Energy. The territory covered 149 towns across Connecticut, and data for modeling were aggregated to a 2-km grid (with 2851 grid cells covering the region). To investigate storm database representativeness in ML model training, we used an events dataset of extratropical storms and thunderstorms, which had distinct characteristics. Extratropical storms last a long time (several hours to days) and exhibit strong sustained winds, rain for a long duration, or both. Thunderstorms are associated with convective events producing lightning, heavy rain, and, sometimes, large hail; they are generally shorter in duration (lasting several minutes to a few hours). Based on the storm characteristics of the outage events in 2005–2017 across Connecticut, we used the method described by Cerrai et al. [8] to classify 92 extratropical storms and 49 thunderstorms of varying severity. This resulted in a total of 262,292 observations for extratropical events and 139,699 for thunderstorm events. The OPM integrated weather variables, utility infrastructure, land cover, vegetation, and historical power outages for each storm event. Details of the data sources follow.

2.1. Weather

The weather data we used in this study represented analyses from the Weather Research and Forecasting model (WRF v.3.7.1), with initial and boundary conditions driven by global data from the National Center for Environmental Prediction (NCEP) Global Forecast System (GFS). We used the high-resolution inner nest (2-km) WRF outputs, which served as the grid to aggregate data on power outages, utility infrastructure, land cover, and vegetation. The input weather variables are summarized in Table 1.

We processed the WRF weather analysis data for each event to extract the following information

- Maximum values during the storm event for selected variables (MaxWind10m, MaxGust, MaxPreRate, MaxSoilMst, MaxTemp, MaxSpecHum, and MAXAbsVor);
- Mean values of selected variables calculated for a 4-h window centered on the highest sustained winds during the event (MeanWind10m, MeanGust, MeanPreRate, MeanSoilMst, and MeanTemp);

- Duration (in hours) of sustained winds at a height of 10 m exceeding 5 m/s and 9 m/s (wgt5 and wgt9);
- Duration (in hours) of wind gusts above 13 m/s (ggt13);
- Continuous duration of sustained winds at a height of 10 m exceeding 5 m/s and 9 m/s (Cowgt5 and Cowgt9).

2.2. Utility Infrastructure

Utility distribution infrastructure contains multiple isolating devices, including electric fuses, reclosers, switches, and transformers. The infrastructure supplies a utility to the customers in the service territory. Eversource Energy provided us with proprietary utility infrastructure data, geographically aggregated to the WRF model's inner domain 2-km grid cells, to provide "SumAssets"—that is, a count of assets per grid cell—for the OPM, as shown in Table 1. "SumAssets" refers to the total number of fuses, reclosers, switchers, and transformers per 2-km grid cell. This variable is typically the most important in a trained OPM and serves as an offset in the model.

2.3. Land Cover

Land cover data with detailed vegetation and urbanization patterns were provided by the University of Connecticut Center for Land Use Education and Research (CLEAR) [15]. Since the interaction of trees next to the overhead lines with the infrastructure during storms is the major cause of outages, we used tree-related land cover variables (that is, the percentages of coniferous forest, deciduous forest, and developed area) per grid cell in the OPM, as shown in Table 1. This process is described in detail by Wanik et al. [6], He et al. [7], and Cerrai et al. [8].

Table 1. Explanatory variables included in the outage prediction model (OPM).

Variable	Description	Unit
wgt5	Duration of wind at 10 m height above 5 m/s	hr
wgt9	Duration of wind at 10 m height above 9 m/s	hr
Cowgt5	Continuous hours of wind above 5 m/s	hr
Cowgt9	Continuous hours of wind above 9 m/s	hr
ggt13	Duration of wind gusts above 13 m/s	hr
MAXWind10m	Maximum wind at 10 m height	m/s
MEANWind10m	Mean wind at 10 m height	m/s
MAXGust	Maximum wind gust	m/s
MEANGust	Mean wind gust	m/s
MAXPreRate	Maximum precipitation rate	mm/hr
MEANPreRate	Mean precipitation rate	mm/hr
MAXSoilMst	Maximum soil moisture	mm/mm
MEANSoilMst	Mean soil moisture	mm/mm
MAXTemp	Maximum temperature	K
MEANTemp	Mean temperature	K
TotPrec	Total accumulated precipitation	mm
MAXSpecHum	Maximum specific humidity	kg/kg
MAXAbsVor	Maximum absolute vorticity	$10^{-5}/s$
PercConif	Percent of coniferous forest	%
PercDecid	Percent of deciduous forest	%
PercDeveloped	Percent of developed area	%
SumAssets	Count of assets	count
LAI	Leaf area index = leaf area / ground area	m^2/m^2

2.4. Vegetation

Since the seasonal variability of the number of leaves on trees is not explained by land cover variables, we used the weekly climatological Leaf Area Index (LAI). We processed this dataset through the quality-controlling algorithm described by Cerrai et al. [8], which used NASA Earth Observations

(NEO) data derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA's Terra and Aqua satellites [16]. We interpolated the 0.1 degrees resolution LAI dataset at the WRF 2-km resolution grid.

2.5. Historical Power Outages

Historical power outages were reported by the Eversource Outage Management System (OMS). Each outage report in the OMS record has geographical coordinates, the number of affected customers, outage start time, power restoration time, device name, outage duration, cause, and weather description. We aggregated the number of power outages during each storm period per 2-km grid cell and utilized the count in the OPM according to geographical coordinates and outage start time.

3. Outage Prediction Model

The OPM devised in this study used multiple tree-based ML regression algorithms (DT, RF, BT, and ENS), which have been used extensively and are described by Wanik et al. [6] and Cerrai et al. [8]. We structured the OPM using the weather analysis (for training) or forecast (for prediction), infrastructure, land cover, and LAI data as the input variables, called predictors X ; these included all information related to power outages. We used actual power outages as the target variable Y . We split the historical event data into training and testing datasets, both containing predictors X and target variable Y . We created ML models relating to X and corresponding Y from the training data and used them to predict target variable Y , given X from the testing data. For different ML regressions, the created models and outputs were different—that is, predicted Y s were different for DT, RF, BT, and ENS with the same predictors X input of the testing data, similar to those in Cerrai et al. [8]. We proposed the use of an optimized model (OPT) that linearly weighted the Y outputs from the four different ML regressions to provide the final OPM prediction value as the power outage prediction for the tested severe storm. The OPM is described next.

The outage prediction output for each storm was given by the OPT, a linear combination of predictions from the DT, RF, BT, and ENS models, based on the formula shown in (1):

$$F_j = CM_j = \sum_{i=1}^4 c_i m_{ij}, \quad (1)$$

where F_j is the power outage prediction for the j -th storm event; $M_j = [m_{1j} \ m_{2j} \ m_{3j} \ m_{4j}]^T$, and m_{1j} , m_{2j} , m_{3j} , m_{4j} represent outage prediction outputs from the DT, RF, BT, and ENS models for the j -th storm event; and $C = [c_1 \ c_2 \ c_3 \ c_4]$, and c_1 , c_2 , c_3 , c_4 represent the weights of DT, RF, BT, and ENS outputs, respectively.

To compute the coefficient C , we optimized the model performance according to the least square errors of predictions F to actuals Y in (2), based on all predicted events:

$$\arg \min_C \|F - Y\|_2 \text{ subject to } C \geq 0, \quad (2)$$

where Y represents actual outage data for all predicted events, and the coefficient C is restricted by nonnegativity, as the outages number is nonnegative.

The coefficient C for each storm is performed by considering only the remaining storms in the out-of-sample validation. Such a coefficient is used for predicting outages related to the excluded storm. The weights for these four models in the OPT model are $RF > ENS > BT > DT$ (i.e., $c_2 > c_4 > c_3 > c_1$).

4. Methods

4.1. Experiment

A flowchart of the evaluation methodology for the storm types (extratropical storms and thunderstorms) is shown in Figure 1. We divided the evaluation we conducted in this study into two parts.

Figure 1. Flowchart of methodology.

First, we trained the OPM with varying numbers of events and quantified the resulting evaluation metrics for both extratropical storms and thunderstorms. We used learning curves for different models (DT, RF, BT, ENS, and OPT) and evaluation metrics to quantify the sample size dependence of OPT error [17]. In ML techniques, the learning curve characterizes the relationship of evaluation metrics and varying training amounts (that is, varying training sample sizes for a given iteration number or different training number of iterations for a given sample size), which shows a measure of predictive performance as a function of a varying number of training samples [18] or iterations [19]. The learning curves are used to search the optimal model performance by increased training sample sizes or iterations until the model performance converges at one given sample size or one iteration number [20]. For regression problems, statistical evaluation metrics are generally used for describing the model performance in response to different sample sizes or numbers of iterations [21]. We used a random sampling method based on a 50 times repeated out-of-sample validation [22]. Specifically, we randomly selected subsets of training events, and run leave-one-storm-out cross-validation [23–25]. This was repeated 50 times to make sure all the events were selected and the predictions were averaged for all the events. This method selected a random subset ranging from 10 to 80 events (in increments of 10) from the 92 extratropical storms and a random subset from 10 to 40 events (in increments of 5) from the 49 thunderstorms as the training dataset. We studied valuation metrics for all events as well as for the top 10 percent in terms of outages, as the latter are the most important to decision makers.

In the second part, we explored the necessity of a training dataset that was representative in terms of outage severity. We divided the training dataset into events of low severity (53 events), moderate severity (31 events), and high severity (8 events) in terms of storm-based outages and included a fourth range comprising all events (92 events) in Table 2. Moderate- and high-severity events can cause outages of significant duration, and utilities may require outside mutual assistance crews to restore power efficiently. Low-severity events can typically be handled by a utility's local crews without asking for mutual assistance. Next, we randomly selected 12 events to explore the potential problem of OPM bias: three with low severity, four with moderate severity, and five with high severity. To characterize and provide a better understanding of these 12 testing events, Table 3 gives a statistical summary of the MAXWind10m attributes (that is, minimum value, maximum value, mean value, standard deviation, and values corresponding to the 10th, 25th, 50th, and 75th percentiles (P10, P25, P50, and P75)).

Table 2. Distribution of four different training datasets.

Training Dataset	Outage Ranges in Training Dataset	Event Frequency in Calibration
Low severity events	[0, 200]	53
Moderate severity events	[200, 1000]	31
High severity events	[1000, 3590]	8
All events	[0, 3590]	92

Table 3. MAXWind10m attribute of 12 tested extratropical storm events.

Tested Events	Actual Outages	Min (m/s)	Max (m/s)	Mean (m/s)	Standard Deviation	P10 (m/s)	P25 (m/s)	P50 (m/s)	P75 (m/s)
Event 1	70	3.76	15.22	6.09	1.24	4.93	5.32	5.87	6.49
Event 2	121	4.05	16.20	6.61	1.19	5.50	5.92	6.41	7.03
Event 3	184	2.85	15.28	5.09	1.27	3.90	4.27	4.82	5.59
Event 4	310	2.85	15.28	5.09	1.27	3.90	4.27	4.82	5.59
Event 5	408	3.94	20.46	6.09	1.43	4.96	5.33	5.79	6.35
Event 6	490	4.14	21.48	6.97	1.41	5.75	6.15	6.69	7.38
Event 7	659	3.35	21.64	6.98	1.94	5.08	5.78	6.72	7.68
Event 8	1141	3.98	20.13	7.51	1.56	6.10	6.65	7.25	7.93
Event 9	1299	3.84	29.07	7.73	2.00	5.74	6.50	7.47	8.47
Event 10	1921	3.56	22.26	6.96	1.65	5.47	6.00	6.65	7.46
Event 11	2918	5.41	23.37	9.00	2.03	7.24	7.80	8.58	9.59
Event 12	3590	5.39	22.54	9.31	2.03	7.51	8.09	8.93	9.92

We defined a severity-conditioned model of sub-setting the training dataset to the predicted event's outage severity. We used cross entropy (CE) [26] to quantify the relatedness of the weather variables' distribution between the events in the training dataset and the tested event. Specifically, we calculated the difference between the weather variables' distribution of the tested event (sample A, p) and this of the mean training dataset (sample B, q), as follows:

$$CE(p, q) = - \sum_{i=1}^N p(x_i) \log q(x_i), \quad (3)$$

where $CE(p, q)$ is the CE of the two discrete distributions (p and q , representing the training and one tested event) of the same parameter. Smaller $CE(p, q)$ indicates that distribution q of the tested event is closer to the mean distribution p of the training events dataset. N is the length-of-steps number (N was equal to 100 in this study); i is one step number; x_i is the weather variable for step i ; $p(x_i)$ is the training events probability of variable x_i ; and $q(x_i)$ is the tested event probability of variable x_i .

The difference between the maximum and minimum of the parameter for sample A and sample B is called D_{pq} . x_i is a range datum, from $D_{pq} * (i - 1)$ to $D_{pq} * i$. $p(x_i)$ equates to the ratio of the frequency

of having the data inside x_i to the total grid number 2851. $q(x_i)$ equates to the ratio of the frequency of having the data inside x_i to the total grid number 2851.

4.2. Evaluation Metrics

To quantify the uncertainty of the OPM with varying sample sizes, we calculated several evaluation metrics for the experiments.

We used absolute error (AE) to measure the difference between the total actual (y_i) and predicted ($\hat{y}_i$) service territory outages from each event i . The AE Q25, AE Q50, and AE Q75 represent the first quartile, second quartile, and third quartile of the sorted absolute error data, respectively. AE is calculated as:

$$AE = |\hat{y}_i - y_i|. \quad (4)$$

We used mean absolute percentage error (MAPE) to measure the mean relative error as a percentage. It is defined as:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \quad (5)$$

We used centered root mean square error (CRMSE) to quantify the random component of the error, which is calculated as follows:

$$CRMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\hat{y}_i - y_i - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)}{n} \right)^2}. \quad (6)$$

We used R-squared (R^2), called the coefficient of determination, to measure the goodness of fit of the various model predictions to the actual outages. R-squared is calculated as follows:

$$R^2 = \frac{\left(\sum_{i=1}^n y_i \hat{y}_i - \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n \hat{y}_i)}{n} \right)^2}{\left(\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \right) \left(\sum_{i=1}^n \hat{y}_i^2 - \frac{(\sum_{i=1}^n \hat{y}_i)^2}{n} \right)}. \quad (7)$$

In addition, we used Nash–Sutcliffe efficiency (NASH), which is the generalized version of R-squared, to determine how well the prediction fit the actual outages. NASH values below zero indicate model performance worse than climatology, while NASH close to 1 indicates accurate prediction. NASH is calculated as follows:

$$NASH = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n \left(y_i - \frac{\sum_{i=1}^n y_i}{n} \right)^2}. \quad (8)$$

For the evaluation metrics, smaller AE, MAPE, and CRMSE mean lower prediction bias, and larger R^2 and NASH mean more advanced model performance.

To demonstrate the sample size dependence of evaluation metrics in terms of MAPE and CRMSE metrics, we calculated the relative change (Δ) of these evaluation metrics as a function of sample size (Equations (9) and (10)):

$$MAPE_I = \frac{MAPE_I - MAPE_\theta}{MAPE_\theta}, \quad (9)$$

$$\Delta CRMSE_I = \frac{CRMSE_I - CRMSE_\theta}{CRMSE_\theta}, \quad (10)$$

where the subscript θ refers to the largest sample size evaluated for each storm type, and I takes on various values for each of the sample sizes we evaluated in the experiment.

5. Results and Discussions

In this section we have discussed a two-part analysis. In part one, we quantified the OPM error metrics associated with different training sample sizes for both extratropical and thunderstorm events. We studied the learning curves for different models (RF, ENS, BT, DT, and OPT) and we evaluated how OPT model metrics and errors changed for different sample sizes. The underfitting and model performance are also discussed in this part.

In the second part, we present the results of sub-setting the OPM's training dataset to historical events with outage severity similar to that of the predicted event. We compared the model performance of the severity-conditioned model and standard model and we showed the bias reduction in predicting low- and high-severity events using representative training datasets in the severity-conditioned model. Finally, we showed CDFs of key weather variables and calculated the cross entropy between the training dataset and tested event data to explain why sub-setting the OPM's training dataset was significant to the OPM accuracy.

5.1. Part One: Quantify the Uncertainty of the OPM Associated with Varying Sample Sizes

This section discusses the results from part one of the study, quantifying the uncertainty of the OPM with varying sample sizes.

The average number of times each storm event was selected during an iteration as the training dataset for the model fitting with different sample sizes is shown in Figure 2 for the two storm types.

Figure 2. Frequency of each storm event used in the training set versus different sample sizes (1—extratropical storms, 2—thunderstorms).

Since moderate- and high-severity events cause substantial power outages (more than 200 in the event period across Connecticut), we studied additional learning curves for the most severe 10% of events for each storm type. Figure 3 shows the MAPE and CRMSE learning curves for all events (panels a b) and the most severe 10% of events (panels c and d) for extratropical storms (blue) and thunderstorms (green), with varying training sample sizes. Note that the OPT model (cross points) had the lowest MAPE and CRMSE learning relative to the RF, ENS, BT, and DT models, and the DT models (triangle points) had the highest for both storm types, except the top 10% of thunderstorm events. The performance of the OPT and RF models improved when the sample sizes increased for both storm types, except for the CRMSE for the top 10% of thunderstorm events. The explanation for these poor CRMSE results is that this category had only five events in the testing dataset, and they were limited and not representative. For increasing sample sizes, the MAPE of the ENS and BT models decreased for the top 10% of severe extratropical events. This decrease was not evident for thunderstorms or

for all extratropical events. As the highest MAPE and CRMSE learning in Figure 3 happened at the sample sizes 10 and 20, these sample sizes are not acceptable for all models. The MAPE and CRMSE learning curves did not converge for either extratropical or thunderstorm events at the maximum sample size, indicating that more events are needed to improve OPM accuracy for both thunderstorm and extratropical storm types.

Figure 3. Learning curves (mean absolute percentage error (MAPE) and centered root mean square error (CRMSE)) for all (panels a and b) and top 10% storm events (panels c and d) versus different sample sizes (1—extratropical storms, 2—thunderstorms).

Since the OPT performed better than the RF, ENS, BT, and DT models, we have focused the rest of the discussion on the performance of this model. Based on the learning curves of the OPT model, it looked like the performance would drop even more if more storms were added. Elaborating on the uncertainty of OPT with varying sample sizes, Tables 4 and 5 summarize the evaluation metrics (AE Q25, AE Q50, AE Q75, MAPE, CRMSE, R-squared, and NASH, described in Section 4) of validation results of the OPT model for different sample sizes for extratropical storms and thunderstorms, respectively.

Table 4. Evaluation metrics of validation results of training random 10 to 80 extratropical storms for the OPT model.

Sample Size	AE q25	AE q50	AE q75	MAPE	CRMSE	R2	NASH
10	98	152	270	135%	435	0.55	0.39
20	76	125	269	116%	379	0.61	0.54
30	58	113	278	98%	347	0.65	0.61
40	55	116	284	93%	311	0.71	0.69
50	63	104	284	103%	317	0.69	0.68
60	42	81	230	79%	271	0.78	0.77
70	34	94	234	81%	253	0.8	0.79
80	45	75	166	68%	199	0.88	0.87

Table 5. Evaluation metrics of validation results of training random 10 to 40 thunderstorms for the OPT model

Sample Size	AE q25	AE q50	AE q75	MAPE	CRMSE	R2	NASH
10	55	123	221	68%	183	0.44	0.41
15	62	121	200	65%	176	0.47	0.45
20	62	111	207	63%	175	0.47	0.46
25	63	113	182	60%	167	0.52	0.51
30	61	113	194	60%	166	0.52	0.52
35	64	98	172	55%	154	0.59	0.58
40	58	90	139	51%	139	0.66	0.66

Table 4 shows, for extratropical events, how AE Q25, AE Q50, AE Q75, MAPE, and CRMSE decreased and R-squared and NASH increased from sample sizes ranging from 10 to 80 events. In terms of evaluation metrics from sample sizes 10 to 80, AE Q25, AE Q50, and AE Q75 decreased by 54, 77, and 104 outages, respectively; MAPE decreased from 135% to 68%; CRMSE decreased by 236 outages; R-squared increased from 0.55 to 0.88; and NASH increased from 0.39 to 0.87. Specifically, at sample size 30, the NASH of the model attained an acceptable value around 0.6, but CRMSE was still high. Performance of the model with sample sizes 10 and 20 was not acceptable. At sample size 60, the CRMSE and MAPE showed a remarkable decrease (more than 50%), and R-squared and NASH showed a remarkable increase (38% and 55%, respectively).

Table 5 shows, for thunderstorms, how AE Q50, AE Q75, MAPE, and CRMSE decreased while R-squared and NASH increased for sample sizes from 10 to 40. In terms of evaluation metrics for sample sizes 10 to 40, AE Q50 and AE Q75 decreased by 33 and 83 outages, respectively; MAPE decreased from 68% to 51%; CRMSE decreased by 44 outages; R-squared increased from 0.44 to 0.66; and NASH increased from 0.41 to 0.66. At sample size 40, AE Q75 and CRMSE showed a remarkable decrease, and R-squared and NASH showed a remarkable increase of more than 0.6. We recommend 40 as the smallest sample size for the OPT of thunderstorm events.

Figures 4 and 5 show OPM validation results of the OPT model based on three different sample sizes (10, 40 and 80) for the extratropical storms and three (10, 25, and 40) for the thunderstorms. In each figure, the vertical and horizontal axes represent the log-scale predicted and actual outages for each event and the maximum and minimum estimates for it, based on the 50 iterations in the top and bottom caps of the error bar. Note that the OPM calibrated with a training dataset of 10 events had the strongest underestimation of high-severity events and the strongest overestimation for low-severity ones for both storm types.

Figure 4. Actual outages vs. predicted outages for a sample size of random 10, 40, and 80 extratropical storms used as training data in the OPM.

Figure 5. Actual outages vs. predicted outages for a sample size of random 10, 25, and 40 thunderstorms used as training data in the OPM.

To explore further the uncertainty of the OPM with varying sample sizes, we studied the learning capacities of the OPT model. The corresponding error changes—that is, ΔMAPE_I and ΔCRMSE_I , calculated by Equations (9) and (10)—are shown in Tables 6 and 7 for extratropical storms and thunderstorms, respectively. Looking first at extratropical storms, note how much ΔMAPE_I and ΔCRMSE_I decreased for all the events, and how much they decreased for the top 10 most severe events at different training sample sizes.

In the comparison between outages predicted by the OPT model and actual outages from all thunderstorms, ΔMAPE_I decreased from 33% to 8% with the increase of the sample size from 10 to 35, and ΔCRMSE_I decreased from 32% to 11%. For the most severe 10% of thunderstorms (five events), ΔMAPE_I decreased from 82% to 18%, and ΔCRMSE_I changed from −19% to −7%. As mentioned, since five events are not representative, and strong thunderstorms are hard to predict, the CRMSE learning curve and ΔCRMSE_I did not show the expected tendency to decrease.

The low sample sizes can cause underfitting and lead to prediction bias for tree-based ML models [27–29]. The sample sizes below 60 for extratropical events and below 40 for thunderstorms show obvious underfitting effects. Using more sample sizes in the training datasets could reduce underfitting when training OPMs.

In summary, from these analyses, we quantified the uncertainty of the OPM with varying sample sizes. These findings may be used widely in the field of outage prediction modeling for selecting useful sample sizes, as well as to validate the new OPM with certain sample sizes. It shows how much evaluation metrics improve and how much uncertainty is lost with an increase in the sample size of the training dataset.

As explained in Section 2, extratropical storms and thunderstorms are different, and Figure 3 shows their learning curves are also different. For the same sample sizes (that is, 10, 20, 30, and 40), the OPM displays lower MAPE and CRMSE errors for extratropical storms than for thunderstorms. In other words, the OPM performs better for extratropical storms than for thunderstorms, which may be attributed to the greater challenge of forecasting convective events [30].

A limitation of this study is that the lower number of thunderstorms (49) relative to the number of extratropical storms (92) yielded a worse model performance. Since thunderstorms occur more often in humid areas during warm summer months while extratropical storms can happen in any season of the year, we have more of the latter in Northeastern United States.

Table 6. $\Delta MAPE_I$ and $\Delta CRMSE_I$ for all and top 10% extratropical events with different sample sizes.

Sample Size I	All Storms $\Delta MAPE_I$	All Storms $\Delta CRMSE_I$	Top 10% Severe Storms $\Delta MAPE_I$	Top 10% Severe Storms $\Delta CRMSE_I$
10	99%	119%	86%	112%
20	71%	90%	76%	91%
30	44%	74%	72%	83%
40	37%	56%	66%	62%
50	51%	59%	48%	64%
60	16%	36%	34%	47%
70	19%	27%	21%	30%
80	0%	0%	0%	0%

Table 7. $\Delta MAPE_I$ and $\Delta CRMSE_I$ for all and top 10% thunderstorms with different sample sizes.

Sample Size I	All Storms $\Delta MAPE_I$	All Storms $\Delta CRMSE_I$	Top 10% Severe Storms $\Delta MAPE_I$	Top 10% Severe Storms $\Delta CRMSE_I$
10	33%	32%	82%	−19%
15	27%	27%	59%	−14%
20	24%	26%	59%	−18%
25	18%	20%	45%	−9%
30	18%	19%	41%	−14%
35	8%	11%	18%	−7%
40	0%	0%	0%	0%

5.2. Part Two: Sub-setting the Training Dataset to Events Representative of the Severity of the Predicted Event

This section illustrates the results from part two of the study, which addressed the issue of underestimation in severe events and overestimation in weak ones. In this part, specifically, we evaluated which model configuration was most suitable for predicting extratropical events in different ranges of outage severity, not always using all the events in the training for a standard approach. As mentioned in Section 4, we selected four representative calibration datasets; their scatter plots of predicted versus actual outages are shown in Figure 6. The two parallel red lines represent 50% OPM overestimation (the top red line) and 50% OPM underestimation (the bottom red line), whereas the black line between them shows the 45-degree line at which the predicted and actual agreed.

Figure 6. Actual vs. predicted outages of 12 holdout tested extratropical storm events by training the model with four different representative calibration datasets [0, 200], [200, 1000], (1000, 3590], and [0, 3590] according to outage severity.

Figure 6 (panel a) shows the results we obtained by using different training datasets for predicting three low-severity events. Three holdout low-severity events were captured by the OPM with the low-outage training dataset [0, 200], while others were mostly overestimated. This was expected, because high-outage storms are most dissimilar to low-outage ones.

Figure 6 (panel b) displays the prediction results for the four moderate-severity events. The OPM with the training dataset of all storms captured all moderate events between the $\pm 50\%$ error bounds and performed better than other OPMs.

Figure 6 (panel c) shows the OPM validation results for the five high-severity events. Four-fifths holdout high-severity events were captured by the OPM with the high-outage training dataset (1000, 3590], performing better than the other OPMs, which underestimated almost all the high-severity events.

To overcome overestimation of low-severity events, the severity-conditioned model used the training dataset of low-severity events, while to overcome underestimation of high-severity events, it used the training dataset of high-severity events. To predict the moderate-severity events, it used the training dataset of all events. To describe better the advantages of the severity-conditioned model, we calculated the MAPE, CRMSE, and NASH of the standard OPM (the 12 black points in Figure 6) and the severity-conditioned model (the blue points in panel a, black points in panel b, and orange points in panel c in Figure 6) for the prediction of 12 events. The standard OPM had a MAPE of 44%, CRMSE of 693, and NASH of 0.44. The severity-conditioned model had a MAPE of 27%, CRMSE of 446, and NASH of 0.84, hence showing more advanced bias reduction than the standard OPM. Thus, conditioning the training dataset to events representative of the predicted event's severity yielded bias reduction.

We believe the aforementioned results highlight the effect of weather features on outage severity [31]. We attribute the systematic errors in the OPM predictions to the problem of representativeness of the training events dataset. Since MAXWind10m and MAXGust are the key

weather variables used by utilities to distinguish and classify the severity of an event, and MAXGust exhibited the best CE performance in identifying similarities between predicted events and events in the training datasets, we chose these two parameters to present similarities and differences between the CDFs (cumulative distribution function) of predicted and training events.

To demonstrate the aforementioned aspect, Figure 7 shows the CDFs of MAXWind10m (panels a, c, e) and MAXGust (panels b, d, f) for storm events and the calibration datasets of the different severity groups. We randomly selected one low-severity event (70 outages on 10 March 2017), one moderate-severity event (408 outages on 30 December 2008), and one high-severity event (1419 outages on 25 October 2008) from the 92 extratropical storms. For the purpose of discussing the relationship of the predicted storm with the calibration dataset, we used purple lines representing the CDF plots of mean MAXWind10m and mean MAXGust of low-severity training events (panels a and b), all-severity training events (panels c and d), and high-severity training events (panels e and f). The MAXWind10m and MAXGust CDFs of the tested low-severity event were similar to the corresponding CDFs of mean MAXWind10m and mean MAXGust of low-severity training events. The MAXWind10m and MAXGust CDFs of the tested moderate-severity event were similar to corresponding CDFs of mean MAXWind10m and mean MAXGust of the all-severity training events. The MAXWind10m and MAXGust CDFs of the tested high-severity event were similar to the corresponding CDFs of mean MAXWind10m and mean MAXGust of the high-severity training events.

Figure 7. Cumulative distribution function (CDF) plots of MAXWind10 (panels a, c, e) and MAXGust (panels b, d, f) for one tested low-severity (70 outages on 10 March 2017), one tested moderate-severity (408 outages on 30 December 2008) and one tested high-severity event (1419 outages on 25 October 2008) and training dataset of low-severity [0, 200], high-severity (1000, 3590) and all-severity events [0, 3590].

Table 8 shows the CE calculation results for MAXWind10m (“W”) and MAXGust (“G”) for three of the aforementioned tested events. The CE of MAXWind10m between the tested low-severity event and the mean low-severity training events was very close to the CE of the mean all-severity training events and lower than the CE of the mean moderate- and mean high-severity training events. The CE of MAXGust between the tested low-severity event and the mean low-severity training events was the lowest relative to the mean moderate-, mean high-, and mean all-training events.

Table 8. Cross entropy of MAXWind10m (W) and MAXGust (G) between tested events and mean MAXWind10m and MAXGust of four training datasets.

Outage Events	Training [0, 200]		Training [200, 1000]		Training (1000, 3590]		All Events	
	W	G	W	G	W	G	W	G
Low-severity event (outages = 70)	3.8	9.1	4.7	11.3	6.3	11.4	3.6	10.9
Moderate-severity event (outages = 408)	4.4	6.2	3.2	5.1	4.1	7.6	3.5	4.5
High-severity event (outages = 1419)	5.9	9.6	4.0	5.4	3.6	3.9	5.0	7.6

The CE calculation results for the tested moderate-severity event (408 outages on 30 December 2008) were as follows: the CE of MAXWind10m between the tested moderate-severity event and the mean all-training events datasets was 3.5 and very close to the lowest CE (3.2) of the mean moderate-training events. The CE of MAXGust between the tested moderate-severity event and the mean all events was the lowest (4.5) relative to the training of the mean low-, mean moderate-, and mean high-severity events.

The CE calculation results for the tested high-severity event (1419 outages on 25 October 2008) were as follows: the CE of MAXGust and MAXWind10m between the tested high-severity event and the mean high-severity training events dataset both had the lowest value relative to the mean low- and mean moderate-severity and the mean all-training events datasets.

The ML models could recognize the severity of the testing weather event when the weather patterns and information of the training weather event were similar to those of the testing event. Specifically, we showed that the CE calculated using the two wind variables could correctly categorize the low and high severity tested events’ in the corresponding category of the training dataset. For the medium severity events we could correctly categorize them in the entire training dataset. These CE results explain the underestimation biases observed when we used the training dataset of low-severity events to test a high-severity event and the overestimation biases observed when we used the training dataset of high-severity events to test a low-severity event. This indicates that the predictive accuracy of the OPM is sensitive to the representativeness of the training weather events dataset.

6. Conclusions

This paper uniquely contributes to the quantification of the uncertainty in outage prediction modeling associated with the sample size and representativeness of its training dataset. Using evaluation metrics and learning curves, we quantified how much the reliability of outage predictions improved, and how much the bias decreased, with increasing sample size. This study benefits the utilities for selecting the number of sample sizes to train an OPM. Based on this study, we could determine the minimum sample size for an acceptable model performance for training an OPM in Eversource Energy’s Connecticut service territory, which is beneficial to the utility to do management. This investigation based on Eversource Energy’s Connecticut service territory may apply to evaluating the uncertainty of outage prediction modeling in other service territories with limited historical datasets. Developing a regional OPM model constitutes one of our avenues for future research.

In addition, this paper introduces a new method to address the underestimation and overestimation biases exhibited by past studies in high-severity and low-severity events, respectively. This involves selecting a representative training dataset for each predicted event to reduce the aforementioned bias in OPM prediction. Specifically, we showed that the prediction of low-severity or high-severity events should use a training dataset of correspondingly low-severity or high-severity events, while prediction of moderate events could be based on a training dataset containing all storm events. Comparing the CDFs and cross entropy of MAXWind10m and MAXGust between the tested events' dataset and the different training datasets provides an explanation as to why the predictive accuracy of the OPM is related to the representativeness of the training weather dataset. Future work will focus on implementing this framework of selecting the training events that best represent the forecasted event's severity in an operational OPM. Improving OPM accuracy based on this method would significantly support utility emergency preparedness efforts before high-impact storm events.

The study was limited in terms of the sample sizes for the thunderstorm events, which affected the OPM performance. We believe that, enriching the thunderstorm events' database, the performance for thunderstorms would improve to the level of performance for extratropical storms and overcome underfitting. The methods for conditioning the training datasets from this study could apply to other ML models, but the expected severities may vary across them. Moreover, the system topology, the sequential trajectory of a storm, and the system-operating conditions may also be significant to the OPM and will be investigated in our future research.

Author Contributions: Conceptualization, F.Y. and E.N.A.; methodology, F.Y., E.N.A., and D.W.W.; software, F.Y.; validation, F.Y.; formal analysis, F.Y.; investigation, F.Y.; resources, D.W.W., D.C., F.Y., and M.A.E.B.; data curation, D.W.W., D.C., F.Y., and M.A.E.B.; writing—original draft preparation, F.Y.; writing—review and editing, E.N.A., F.Y., D.W.W., D.C., and M.A.E.B.; supervision, E.N.A.; project administration, E.N.A.; funding acquisition, E.N.A.. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Eversource Energy and DTN LLC.

Acknowledgments: The authors of this publication had research support from Eversource Energy and DTN LLC. In addition, E.N. Anagnostou and D. Cerrai hold stock in ACW Analytics. This publication partially uses classified datasets of the electric grid. We have full access to all of the data in this study, and we take complete responsibility for their integrity and the accuracy of the data analysis.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Executive Office of the President. Economic Benefits of Increasing Electric Grid Resilience to Weather Outage. August 2013. Available online: https://www.energy.gov/sites/prod/files/2013/08/f2/Grid%20Resiliency%20Report_FINAL.pdf (accessed on 17 February 2020).
2. Han, S.R.; Guikema, S.D.; Quiring, S.M. Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models. *Risk Anal. Int. J.* **2009**, *29*, 1443–1453. [CrossRef] [PubMed]
3. Han, S.-R.; Guikema, S.D.; Quiring, S.M.; Lee, K.-H.; Rosowsky, D.; Davidson, R.A. Estimating the spatial distribution of power outages during hurricanes in the Gulf coast region. *Reliabil. Eng. Syst. Saf.* **2009**, *94*, 199–210. [CrossRef]
4. Nateghi, R.; Guikema, S.; Quiring, S.M. Power outage estimation for tropical cyclones: Improved accuracy with simpler models. *Risk Anal.* **2014**, *34*, 1069–1078. [CrossRef] [PubMed]
5. Guikema, S.D.; Nateghi, R.; Quiring, S.M.; Staid, A.; Reilly, A.C.; Gao, M. Predicting hurricane power outages to support storm response planning. *IEEE Access* **2014**, *2*, 1364–1373. [CrossRef]
6. Wanik, D.; Anagnostou, E.; Hartman, B.; Frediani, M.; Astitha, M. Storm outage modeling for an electric distribution network in Northeastern USA. *Nat. Hazards* **2015**, *79*, 1359–1384. [CrossRef]
7. He, J.; Wanik, D.W.; Hartman, B.M.; Anagnostou, E.N.; Astitha, M.; Frediani, M.E. Nonparametric Tree-Based Predictive Modeling of Storm Outages on an Electric Distribution Network. *Risk Anal.* **2017**, *37*, 441–458. [CrossRef]

8. Cerrai, D.; Wanik, D.W.; Bhuiyan, M.A.E.; Zhang, X.; Yang, J.; Frediani, M.E.; Anagnostou, E.N. Predicting Storm Outages Through New Representations of Weather and Vegetation. *IEEE Access* **2019**, *7*, 29639–29654. [\[CrossRef\]](#)
9. Figueroa, R.L.; Zeng-Treitler, Q.; Kandula, S.; Ngo, L.H. Predicting sample size required for classification performance. *BMC Med. Inf. Decis. Mak.* **2012**, *12*, 8. [\[CrossRef\]](#)
10. Nuchitprasittichai, A.; Cremaschi, S. An algorithm to determine sample sizes for optimization with artificial neural networks. *AIChE J.* **2013**, *59*, 805–812. [\[CrossRef\]](#)
11. Tang, X.-S.; Li, D.-Q.; Cao, Z.-J.; Phoon, K.-K. Impact of sample size on geotechnical probabilistic model identification. *Comput. Geotech.* **2017**, *87*, 229–240. [\[CrossRef\]](#)
12. van Proosdij, A.S.; Sosef, M.S.; Wieringa, J.J.; Raes, N. Minimum required number of specimen records to develop accurate species distribution models. *Ecography* **2016**, *39*, 542–552. [\[CrossRef\]](#)
13. Shashaani, S.; Guikema, S.D.; Zhai, C.; Pino, J.V.; Quiring, S.M. Multi-Stage Prediction for Zero-Inflated Hurricane Induced Power Outages. *IEEE Access* **2018**, *6*, 62432–62449. [\[CrossRef\]](#)
14. Bhuiyan, M.A.E.; Nikolopoulos, E.I.; Anagnostou, E.N.; Quintana-Seguí, P.; Barella-Ortiz, A. A nonparametric statistical technique for combining global precipitation datasets: Development and hydrological evaluation over the Iberian Peninsula. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 1371. [\[CrossRef\]](#)
15. University of Connecticut Center for Land Use Education and Research (CLEAR), Connecticut’s Changing Landscape. Available online: <http://clear.uconn.edu/projects/landscape/> (accessed on 17 February 2020).
16. NASA EARTH OBSERVATIONS, LEAF AREA INDEX (8 DAY—TERRA/MODIS). Available online: https://neo.sci.gsfc.nasa.gov/view.php?datasetId=MOD15A2_E_LAI&date=2017-02-01 (accessed on 17 February 2020).
17. Perlich, C. Learning curves in machine learning. In *Encyclopedia of Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 577–580.
18. Samala, R.K.; Chan, H.-P.; Hadjiiski, L.; Helvie, M.A.; Richter, C.D.; Cha, K.H. Breast Cancer Diagnosis in Digital Breast Tomosynthesis: Effects of Training Sample Size on Multi-Stage Transfer Learning using Deep Neural Nets. *IEEE Trans. Med. Imaging* **2019**, *38*, 686–696. [\[CrossRef\]](#)
19. Kim, S.-J.; Giannakis, G.B. An online convex optimization approach to real-time energy pricing for demand response. *IEEE Trans. Smart Grid* **2017**, *8*, 2784–2793. [\[CrossRef\]](#)
20. Perlich, C.; Provost, F.; Simonoff, J.S. Tree induction vs. logistic regression: A learning-curve analysis. *J. Mach. Learn. Res.* **2003**, *4*, 211–255.
21. Vandael, S.; Claessens, B.; Ernst, D.; Holvoet, T.; Deconinck, G. Reinforcement learning of heuristic EV fleet charging in a day-ahead electricity market. *IEEE Trans. Smart Grid* **2015**, *6*, 1795–1805. [\[CrossRef\]](#)
22. Ng, W.; Dash, M. An evaluation of progressive sampling for imbalanced data sets. In Proceedings of the Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on Data Mining, Hong Kong, China, 18–22 December 2006; IEEE: Piscataway, NJ, USA, 2006; pp. 657–661.
23. He, K.; Zha, R.; Wu, J.; Lai, K.K. Multivariate EMD-based modeling and forecasting of crude oil price. *Sustainability* **2016**, *8*, 387. [\[CrossRef\]](#)
24. Kearns, M.; Ron, D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Comput.* **1999**, *11*, 1427–1453. [\[CrossRef\]](#)
25. Cawley, G.C. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In Proceedings of the 2006 IEEE International Joint Conference on Neural Network Proceedings, Vancouver, BC, Canada, 16–21 July 2006; IEEE: Piscataway, NJ, USA, 2006; pp. 1661–1668.
26. Abbas, A.E.; Cadenbach, H.A.; Salimi, E. A Kullback–Leibler View of Maximum Entropy and Maximum Log-Probability Methods. *Entropy* **2017**, *19*, 232. [\[CrossRef\]](#)
27. Gu, Y.; Wylie, B.K.; Boyte, S.P.; Picotte, J.; Howard, D.M.; Smith, K.; Nelson, K.J. An optimal sample data usage strategy to minimize overfitting and underfitting effects in regression tree models based on remotely-sensed data. *Remote Sens.* **2016**, *8*, 943. [\[CrossRef\]](#)
28. Yu, L.; Lai, K.K.; Wang, S.; Huang, W. A bias-variance-complexity trade-off framework for complex system modeling. In Proceedings of the International Conference on Computational Science and Its Applications, Glasgow, UK, 8–11 May 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 518–527.
29. Cucker, F.; Smale, S. Best choices for regularization parameters in learning theory: On the bias-variance problem. *Found. Comput. Math.* **2002**, *2*, 413–428. [\[CrossRef\]](#)

30. Doswell, C. *Severe Convective Storms*; Springer: Berlin/Heidelberg, Germany, 2015.
31. Chen, P.-C.; Kezunovic, M. Fuzzy logic approach to predictive risk analysis in distribution outage management. *IEEE Trans. Smart Grid* **2016**, *7*, 2827–2836. [[CrossRef](#)]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).