Nonparametric Tree-based Predictive Modeling of Storm Outages on an Electric Distribution Network

October 2015

Abstract

This paper compares two nonparametric tree-based models, quantile regression forests (QRF) and Bayesian additive regression trees (BART), for predicting storm outages on an electric distribution network in Connecticut, USA. We evaluated point estimates and prediction intervals of outage predictions for both models using high-resolution weather, infrastructure, and land use data for 89 storm events (including hurricanes, blizzards and thunderstorms). We found that QRF produced better results for high spatial resolutions, while BART predictions aggregated to coarser resolutions more effectively, which would allow for a utility to make better decisions about allocating pre-storm resources. We also found that the predictive accuracy was dependent on the season (e.g. tree-leaf condition, storm characteristics), and that the predictions were most accurate for winter storms. Given the comparable performance characteristics, we suggest that BART and QRF be implemented together to show the complete picture of a storm's potential impact on the electric distribution network.

KEY WORDS: Weather hazards; electric distribution network; quantile regression forests; Bayesian additive regression trees; critical infrastructure outage modeling.

1 INTRODUCTION

Severe weather is among the major causes of damage to electric distribution networks and resultant power outages in the United States [(1)]. In addition to hurricanes, for which significant research has been focused, other more frequent weather systems (e.g. thunderstorms and frontal systems) have caused power outages in Connecticut lasting from several hours up to several days. Accurate prediction of the number of outages associated with storms would allow utility companies to restore power faster by allocating resources more efficiently. Furthermore, to effectively use these outage predictions in decision-making, models must exhibit acceptable accuracy in the spatial distribution of estimated outages and characterization of the prediction uncertainty.

A wide range of models have been employed in hurricane outage modeling, beginning with parametric statistical models. Generalized linear models (GLMs) were utilized by Liu *et al.* [(2)] using negative binomial regression with binary index variables representing storm similarity characteristics. Guikema *et al.* [(3)] explored the effects of tree trimming on hurricane outages with a GLM and a generalized linear mixed model (GLMM). Liu *et al.* [(4)] attempted the use of spatial GLMM for better inference on variables, but did not achieve improved prediction accuracies using random effects or spatial correlation modeling. Han *et al.* [(5)] suggested using more informative descriptive variables with GLM and performed principal components analysis (PCA) as a treatment to transform correlated variables and obtain stable parameter estimates.

Nonparametric models for hurricane outage prediction gained popularity shortly thereafter. Guikema *et al.* [(6)] compared multiple models including classification and regression trees (CART), generalized additive models (GAM), Bayesian additive regression trees (BART) and GLM, and discussed the advantages of nonparametric models over parametric models for outage prediction for hurricanes. Nateghi *et al.* [(7)] expanded the topic to outage duration modeling and concluded that multivariate adaptive regression splines (MARS) and BART had better results than traditional survival analysis models and CART, and that BART produced the lowest prediction error. Guikema *et al.* [8] proposed a two-stage model using classification trees and logistic regression to deal with zero-inflation and GAM for over-dispersion, which helped balance the statistical assumption and prediction. Recently, Nateghi *et al.* [9] highlighted the modifiable areal unit problem (MAUP) and compared predictions from random forests (RF) and BART, concluding that RF benefited from its distribution-free setting and performed the best in outage duration prediction. Among all nonparametric models, tree-based models, and especially multiple trees or forests models, have been used widely and have been generally preferred in modeling hurricane outages.

Aiming to assist the largest utility company in Connecticut, Eversource Energy, in pre-storm decision making, we investigated two models and compared their predictions of spatial outage patterns and their ability to perform statistical inference. This study builds on our previous research that investigated the use of different model forcing data and methods for predicting power outages in Connecticut (Wanik *et al.* [10]). Prediction intervals of model estimates are as important for risk management as point estimations of storm outages; a point estimate only provides a single value at each location to describe the predicted storm outages, while prediction intervals provide a characterization of the uncertainty associated with the prediction. The lack of uncertainty characterization can affect the complex socioeconomic aspects of emergency response. In this paper, we compare two nonparametric tree-based models for the prediction of storm outages on the electric distribution network of Eversource Energy. In keeping with the most recent research in hurricane outage prediction, we selected quantile regression forests (QRF) and Bayesian additive regression trees (BART) as our candidate models because they were capable of both point estimation and prediction interval construction. BART was shown by Guikema *et al.* [6] to be the most accurate of the different hurricane outage prediction models evaluated in that study. QRF is derived from the random forest model, which has been demonstrated by Nateghi *et al.* [9] to provide robust spatial data aggregation and better power outage duration estimates than BART in terms of prediction error.

We seek to address the following three questions about these two models: 1) How accurate are these models in providing the point estimates (single predicted value per storm) of outages for storms of varying severity?; 2) How efficient are these models in evaluating the prediction uncertainty (i.e. the prediction interval)?; 3) Are these models able to predict outages for different spatial resolutions via aggregation?

2 STUDY AREA AND DATA DESCRIPTION

The analysis was performed on a dataset of 89 storms of multiple temporal and spatial scales (i.e. deep and shallow convective events, hurricanes, blizzards and thunderstorms) that occurred during a ten-year period (2005-2014). We selected the explanatory variables based on their potential contribution to outages on the overhead lines when interactions of overhead lines and vegetation occur. All data including distribution system infrastructure, and land cover information (Table 1) were processed on a high-resolution gridded domain (grid spacing: 2x2 km^2) to represent the average conditions in the corresponding Numerical Weather Prediction (NWP) model grid spacing. Further, a seasonal categorization variable was created for each of the 89 storms (Table I) to represent the actual tree-leaf conditions (e.g. leaf-on, leaf-off or transition) at the time of each storm. The study area was the Connecticut service territory of Eversource Energy (Figure 1), which spans 149 towns in Connecticut and is organized into four divisions (central, west, east and south). The outage predictions were analyzed over the corresponding NWP model grid cells, and subsequently spatially aggregated into coarser resolutions (town, division and territory) to investigate the effects of multiple scales.

2.1 Storm Outage Data

An outage is defined as a location where a two-man restoration crew needs to be sent for manual intervention to restore power. Storm outage records were acquired from the utility's outage management system (OMS) and to improve data quality, duplicate outage records and records with cause codes irrelevant to weather (e.g. vandalism or vehicular outage) were deleted from the data. Outages were recorded at the location of the nearest upstream isolating device (i.e. fuses, reclosers, switches, transformers) from where the damage on the overhead line occurred, which may be different from where the actual outage occurred. We made no differentiation of the different outage types to the overhead lines (i.e. a tree leaning on a conductor, a malfunctioning isolating device, or a snapped pole, etc.) because such data were not available to us.

2.2 Weather Simulation and Verification

The Weather Research and Forecasting Model (WRF; Skamarock *et al.* [11]) was devised to simulate the 89 storm events used in our study. The WRF model simulations were initialized and constrained at the model boundaries using NCEP Global Forecast System (GFS) analysis fields [12]. The NWP model is set up with three nested domains with 18, 6 and 2-km of increasing grid spacing (Figure 2). The simulated meteorological variables were summarized into maximum and mean values (Table I). The wind-related variables in the NWP model included wind at 10m, gust winds, and wind stress. The precipitation-related variables comprised of total accumulated precipitation, the precipitation rate, and snow water equivalent (SWE). The mean values of the selected meteorological variables represent the lasting impact of a storm. The means were calculated over the 4-hour window defined by the simulated wind speed, to which hereafter we refer to as the sustained period of the storm. This period is defined by the highest averaged value in the 4-hour running window across the NWP simulation length. The wind-based sustained period was then used to calculate the mean of the other meteorological variables. The maximum values of the meteorological variables represent the peak severity that occurred during the storm; they correspond to the nominal variable value at the time of highest simulated wind speed. Complementing the mean and maximum variables, the duration of winds and gusts above defined thresholds

(9m/s for wind, 18m/s and 27m/s for gust) were used to relate the duration of damaging winds to outages (Table I).

The NWP model simulations were verified by comparing the sustained wind speed (previously defined) for three major events (Hurricane Irene in August 2011; Hurricane Sandy in October 2012; Blizzard Nemo in February 2013) against METAR observations (airport meteorological station data) provided by the National Centers for Environmental Prediction (NCEP) ADP Global Upper Air and Surface Weather Observations [13]. Though not shown here, the NWP model simulations showed acceptable agreement with the airport station data (e.g. low mean bias, and high correlation between actual and simulated sustained wind speed). The reader may refer to Wanik *et al.* [10] for more details on this verification exercise.

2.3 Seasonal Categorization

Storms affecting the distribution network can have a wide range of weather attributes (e.g. heavy snow or rain) that interact with overlying vegetation, and can have differing impact on the outage magnitude and frequency depending on the tree-leaf condition. For example, high winds usually have a greater impact on trees with leaves due to increased wind loading [14], [15]. To capture this dynamic, we grouped our data by season (Table I), which resulted in separate fits for each of the three different seasonal categories. Of the 89 storm events in our dataset, there were 38 storms and 1 hurricane (Irene) during the summer (leaves on) months (June to September); there were 24 storms and 1 hurricane (Sandy) during the spring and fall (transition) months (October, November, April and May); and there were 25 storms during the winter (leaves off) months (December to March).

2.4 Infrastructure and Land Use

The same infrastructure and land use data from Wanik *et al.* [10] study was used in this paper. The sum of isolating devices (e.g. sum of fuses, reclosers, switches, transformers) in each 2-km grid cell was an important predictor in our models, which we attribute to the outage recording methodology (Section 2.1); if only one outage can be recorded at an isolating device, a grid cell with more isolating devices has more chances to record more outages than a grid cell with less isolating devices. Given that outages were recorded at the nearest isolating device and not the actual outage location, the different types of isolating devices were summed up into a single variable ("sumAssets") instead of modeling outages by isolating device type. This variable sets the upper limit on the number of outages that could occur in a 2-km grid cell.

Accurate tree-specific data (i.e. height, species, and health) around overhead lines are difficult to acquire, so we used land cover data aggregated around the overhead lines as a surrogate for the actual tree data. This aggregation differs with previous research (Quiring *et al.* [16]) that used the percentage of all land cover types in a grid cell, regardless of whether or not certain land cover types in that grid cell interacted with the overhead lines (i.e. a waterbody that is in the grid cell but is not close enough to the overhead lines to cause influence). Land cover data (30m resolution) were obtained from the University of Connecticut Center for Land Use Education and Research (CLEAR) [17] and were used to generate percentages of land use per grid cell. Details about the calculation of land use are available in Wanik *et al.* [10].

3 METHODOLOGY

As two systematically different examples of nonparametric tree-based models, QRF and BART utilize different assumptions and techniques for their application. We briefly introduce the benefits and known issues of these two models, followed by measurements and methods for analysis and comparison of the models.

3.1 Quantile Regression Forests

Based on the well-known random forests algorithm by Breiman [18], Meinshausen [19] created the quantile regression forests (QRF) with the idea of quantile regression from econometrics. Similar to the weighted average of all the trees for predicted expected value of response, QRF utilizes the same weights to calculate the empirical distribution function:

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x) 1_{\{Y_i \leq y\}}. \tag{1}$$

The algorithm of QRF can be summarized as:

1. Grow k trees T_t, $t = 1, ..., k$, as in random forests. However, for every terminal node of every tree, take note of all observations, not just their average.

2. For a given $X = x$, drop x down all trees. Compute the weight $w_i(x, T_t)$ of observation $i = 1, ..., n$ for every tree. Compute weight $w_i(x)$ for every observation $i = 1, ..., n$ as an average over $w_i(x, T_t)$, $t = 1, ..., k$. These weight calculations are the same as in random forests.

3. Compute the estimate of the distribution function as in (Equation 1), using the weights from Step 2.

Given the estimated empirical distribution, quantiles and percentiles are readily available. In this case, mean from the estimated distribution is a natural point estimate, which is exactly the same as random forests. Meinshausen [20] has provided an R package called "quantregForest", with dependency on the "randomForest" package by Liaw *et al.* [21]. Our analysis was based on a slightly modified version of "quantregForest" package providing RF prediction as well as desired quantiles.

QRF has already been used in several aspects of natural phenomena, but have not been implemented in storm outage prediction directly. Juban *et al.* [22] tested QRF in approximating the kernel density of short-term wind power, indicating comparatively wide prediction intervals by QRF. Francke *et al.* [23] addressed the better performance of QRF over GLMs in flood-based analysis of high-magnitude sediment transport. Zimmermann *et al.* [24] also took advantage of QRF to study erosion in rainforests.

Suppose we need to generate a prediction interval for a storm event by aggregation. Accurate empirical distributions are highly preferred, but demand a large number of observations. For example, to generate percentiles (actually 101 quantiles including maximum and minimum), we prefer more than 100 observations in each terminal node. However, enforcing too many observations in terminal nodes could introduce bias if the sample size is not large enough. In this study, we compared QRF to BART in order to get a deeper understanding about the importance of prediction intervals in characterizing model performance.

3.2 Bayesian Additive Regression Trees

Bayesian additive regression trees model (BART), introduced by Chipman, George and McCulloch [25, 26], is a high performance derivation of Bayesian classification and regression trees model (CART). It takes advantage of a backfitting MCMC algorithm [27] in generating the posterior sample of CART. Instead of a single regression tree (the mode of posterior tree sample), a sum of regression trees is utilized to estimate the response under normal assumption:

$$Y = \underbrace{\sum_{j=1}^{m} g(x; T_j, M_j)}_{Mean\ Model} + \underbrace{\epsilon, \epsilon \sim N(0, \sigma^2)}_{Variance\ Model}. \tag{2}$$

Here T_j stands for the j^{th} regression tree; M_j stands for the j^{th} set of terminal nodes; m stands for total number of trees. The prior for probability of splitting node η (depth=d_η), which is also presented by Chipman, George and McCulloch [28]:

$$p_{split} = \alpha(1 + d_\eta)^{-\beta}, 0 \leq \alpha \leq 1 \ and \ \beta \geq 0. \tag{3}$$

Similar to Friedman's gradient boosting [29], each terminal nodes μ_{ij} is determined by

$N(0, \sigma_\mu^2)$, where $\sigma_\mu = 0.5/k\sqrt{m}$. An inverse chi-square distribution is set as the prior of σ^2, parameterized with ν and $P(\sigma < \hat{\sigma}) = q$. All of these hyper parameters can be optimized via cross-validation.

Chipman *et al.* [30] provided an R package "BayesTree" based on C and Fortran, with their original work. Pratola *et al.* [31] offered a standalone C++ implementation with fast parallel computation. Kapelner and Bleich *et al.* [32] made the R package "bartMachine" based on rJava, including features like parallel cross-validation and interaction detection, which we used in this paper.

BART has been widely used in risk analysis and the prediction of natural hazards. Guikema *et al.* [6] conducted a comparison of multiple models for estimating the number of damaged poles during storms, and concluded that BART and an ensemble model with BART outperformed other parametric regression methods. Nateghi *et al.* [7] compared BART with traditional survival models in predicting power outage durations in Hurricane Ivan, 2004, and concluded that BART had better performance over parametric survival models. Blattenberger *et al.* [33] implemented BART in predicting binary response of avalanches crossing streets. They compared BART classification with linear and logistic regressions by altering the cutoff probabilities and concluded that BART excelled in predicting binary response.

As a well-defined Bayesian statistical model, BART naturally offers prediction intervals under its model assumptions, but the error term can be misspecified with respect to storm outage modeling. Both modeling the number of outages [6] and outage durations [9] involve errors that do not necessarily follow a normal distribution. In our study, the response variable (the number of outages) seemed to follow a Poisson distribution in grid cells and towns, while a zero-truncated normal distribution seemed to fit better in divisions and the territory (Figure 3; hurricanes are excluded for extreme values.). That is, these errors could approximately satisfy normality and homogeneity of variance in some situations, while the distribution of data aggregated with different spatial resolution can vary greatly. This issue was first discovered by Gehlke and Biehl [34] and later discussed in details by Openshaw [35].

To understand the impact in our study, we would like to study the prediction intervals given by BART in more detail.

3.3 Metrics of Model Performance Evaluation

We will compare the two models using the following metrics:

Mean absolute error (MAE) is an absolute measurement of the point estimate error of n predictions, which is calculated by

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|. \tag{4}$$

Root mean square error (RMSE) measures the magnitude of error as well as its variability, which is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}. \tag{5}$$

MAE is less than or equal to RMSE. Closer difference between MAE and RMSE indicates smaller variability in the point estimate of error. A combination of MAE and RMSE is a common tool in model comparison with respect to point estimates.

Relative error (RE) is also known as the relative percentage error, which is computed by the following normalized average:

$$RE = \frac{\hat{y}_i - y_i}{y_i}. \tag{6}$$

RE is useful for diagnostics of over-prediction or under-prediction, typically offering an indication of bias.

Nash-Sutcliffe efficiency (NSE) is the generalized version of R-squared from parametric regression, and is widely used in hydrology. NSE was introduced by Nash and Sutcliffe [36] and summarized by Moriasi *et al.* [37]. It is calculated by the following:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \overset{unbiased}{=} 1 - \frac{\hat{Var}(error)}{\hat{Var}(y)}. \quad (7)$$

R^2 is intuitively known as "percent variance explained", and pseudo R^2 is a built-in statistic of both "randomForest" package and "bartMachine" package in R. However, since R^2 is always measured with in-sample data, we use the name of NSE to highlight its capability in validation. Without bias and overfitting, NSE is a powerful tool measuring predictions of spatial variability for nonparametric models; NSE values range from negative infinity to 1. For example, NSE for a single storm validation could be negative, while the average NSE for the validation of all the storms remains positive, indicating that predictions for this specific storm is worse than a mean-only model in terms of spatial variability. When calculating NSE, we corrected the bias of total predicted value of each storm by scaling in order to focus on the spatial variability of each storm event.

When R^2 is positive, it is weakly increasing as p/n increases, where p stands for the number of predictors and n stands for number of predictions. In our case, we aggregate predicted values from high resolution to low resolutions, which actually decreases n with predictors fixed. When NSE is positive (<1) for grid cells, we can expect NSE to be positive and even closer to 1 for towns and divisions, since the pseudo p/n increases. Similarly, negative NSE in high resolution may result in even smaller NSE in low resolutions. Thus we may observe a "polarization" effect after aggregation.

Uncertainty ratio (UR) is a benchmark statistic of prediction uncertainty. Denote the prediction interval as $(Q_{lower_i}, Q_{upper_i})$. Similar to the UR used in Özkaynak *et al.* [38], our definition of UR is a generalized version for asymmetric intervals:

$$UR = \frac{\sum_{i=1}^{n}(Q_{upper_i} - Q_{lower_i})}{\sum_{i=1}^{n} y_i}. \quad (8)$$

While UR is computed by summing up all the ranges of prediction intervals, the formal calculation of a prediction interval for each storm is done by summing up the simulated sample of prediction for each grid cell. Larger UR indicates relatively wider ranges of the prediction intervals (which may be less useful than narrower intervals, as they provide less detailed information).

Exceedance probability (EP) is a measurement of the probability that actual value will exceed the prediction interval. In this paper, we calculate the EP for each storm by

$$\hat{P}\{exceed.\} = 1 - \frac{1}{n}\sum_{i=1}^{n} 1_{\{Q_{lower_i} < y_i < Q_{upper_i}\}}. \tag{9}$$

Similar to UR, large EP is unfavorable, and it implies a large chance to have actual values outside the prediction intervals.

Coverage percentage is the opposite of exceedance probability and is defined as

$$Coverage\ \% = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Q_{lower_i} < y_i < Q_{upper_i}\}} * 100\%. \tag{10}$$

In contrast to exceedance probability, we pursue high coverage rate of prediction intervals on actual values.

Rank histogram: A rank histogram provides a diagnosis of bias and dispersion in ensemble predictions (detailed interpretation of this plot can be found in Hamill [(39)]). Suppose we have m predictions for each observation y_i, denoted as $\{\hat{y_{ij}}\}_{j=1,2,\ldots,m}$, and then the "ensemble" prediction is $\hat{y_i} = \overline{\hat{y_{ij}}}$. A good model implies that the response is a realization of the prediction distribution, namely:

$$E\{P[y_{i,\hat{j}-1} < y_i < \hat{y_{i,j}}]\} = \frac{1}{m+1}. \tag{11}$$

A rank histogram could be generated by collecting all the ranks of actual value in their prediction samples, denoted by $\mathbf{R} = (r_1, r_2, \ldots, r_{m+1})$, where r_j is the following average over all the i's:

$$r_j = \overline{\hat{P}\{y_{i,\hat{j}-1} < y_i < \hat{y_{i,j}}\}}. \tag{12}$$

An ideal rank histogram should display a uniform distribution. A rank histogram with a convex function indicates under-dispersive predictions; a concave function implies over-dispersion. A rank histogram with larger values on the right than on the left addresses negative bias of predictions, while positively biased predictions yield large value on the left. In short, rank histogram above average means the distribution of actual values is "denser" than the distribution of predictions and vice versa.

3.4 Methods of Model Performance Comparison

Our analysis is based on dataset containing 89 storms that occurred in the Eversource (Connecticut) service territory, which resulted in 253,739 observations (89 storms with 2,851 grid cells per storm) for model training and validation. Within each storm, we randomly selected two thirds of the observations (n = 1,989) per storm for model training, and training data were grouped by season. The rest of the data (n = 952 grid cells per storm) were used for model validation. Only model validation results will be presented in our model verification statistics.

For the QRF model, we specified a random forest of 1,000 trees (default setting for the "quantregForest" package [(20)]) and 200 minimal instances for each terminal node to generate percentiles. Quantile regression was introduced to obtain 101 percentiles (including minimum and maximum) and predicted empirical distribution for each grid cell. The mean of the predicted empirical distribution (the same as random forest predictions), are recorded as point estimates at the grid cell level. We calculated 80% prediction intervals for each grid cell with the 10% and 90% prediction quantiles from QRF. We then sampled from the predicted empirical distribution 10,000 times per grid cell and aggregated these prediction samples to get empirical distributions in town, division and territory resolution. The weights obtained from step 2 in Section 3.1, were normalized as probabilities to draw prediction sample from training data responses. The mean and quantiles are consistent with predicted distribution, proved by Bickel *et al.* [40]. After that, sample means and 80% intervals are calculated for these 3 granularities. In the end, we combined these point estimations (means) and intervals for different seasons and generated plots of statistics, such as NSE, UR and rank histogram.

For the BART model, a 5-fold cross-validation indicated the following settings for the parameters: 50 trees, $k = 2$, $q = 0.99$, $\nu = 3$, while other parameters remained default. In order to reach the convergence of MCMC, we performed 10,000 burn-in iterations, which produced a momentum sample (discarded). After that, we ran another 10,000 iterations to get the prediction sample of the model, which was used for prediction and validation. Point estimations and prediction samples for each observation in the testing dataset were computed in the Bayesian way. That is, sampling from the mean model (Equation 2) posterior sample and variance model (Equation 2) posterior sample under model assumptions. In contrast to QRF, BART generated prediction samples from a well-defined distribution instead of empirical distribution. Prediction intervals were calculated for BART in a way similar to QRF: the 10% and 90% quantiles from prediction samples for grid cells, or aggregated prediction samples for resolution lower than the grid cell.

We plotted statistics grouped by season and varying spatial resolutions (grid cell, town, division and territory) for evaluation. By inspecting the plots, we intuitively summarized the different behaviors of QRF and BART, followed by a discussion of the observed patterns and their causes.

4 RESULTS AND DISCUSSION

First, we will discuss the QRF and BART performance for predicting Hurricane Irene (2011) and Hurricane Sandy (2012) outages. Then, we will evaluate the consistency and prediction intervals of QRF and BART for different types of storms in our database. This evaluation is based on the statistical metrics described above, evaluated at different spatial resolutions ranging from the 2-km grid cell to town and regional averages.

4.1 Hurricane Outage Modeling

4.1.1 Point Estimate Results

With respect to hurricanes, Table II shows that both QRF and BART performed well in terms of point estimates, compared with a mean model (assuming uniform outages across grid cells). The mean model performed well in predicting total number of outages for Hurricane Sandy, because the randomly selected validation partition happened to capture two thirds of total outages in this case. However, the town level MAE and RMSE reveal that the mean model did not spatially predict the actual outages for Irene and Sandy. Both QRF and BART exhibited small MAE values (5.50 - 8.86 outages per town) with RMSE values close to MAE values, which indicates moderate variance of the point estimates and a lack of large residuals. For these tropical storm cases, BART showed less error than QRF in town resolution but did not exhibit an overwhelming advantage. QRF predictions were slightly more spread-out than BART as shown in the scatter plots of Figure 4, which were consistent to the error metrics in Table II. Figure 5 illustrates the similar capability of QRF and BART in explaining the spatial variability of the predictions using the town-aggregated estimates; both QRF and BART predicted that the majority of outages from Irene and Sandy would be in central and southwestern Connecticut.

4.1.2 Prediction Interval Results

Both models exhibited different characteristics in terms of their prediction intervals. Figure 4 shows that QRF produced more conservative town-resolution predictions by offering longer prediction intervals and a higher coverage rate than BART; coverage values for Irene and Sandy were 84% (69%) and 87% (77%) for QRF (BART) using the 80% confidence intervals. However, BART had narrower prediction intervals and was able to the cover actual number of outages for both Irene and Sandy in Table II, while QRF failed to cover Irene's actual number of outages in its interval. We also noticed that BART produced symmetric intervals from the normal distribution, while QRF generated asymmetric intervals from the empirical distribution. Although we observed comparatively good results for BART, we were unable to conclude which model was superior in terms of predicting storm outages from analyzing these two hurricanes alone. Further investigation of the model performance for both hurricanes and the remaining less severe weather events follows next.

4.2 Comparison of Model Performance for All Storm Events

4.2.1 Point Estimate Results

In this section, we will examine how QRF and BART explained the magnitude and spatial variation of outages (i.e. point estimates), and also examine how both models' prediction intervals explained the variability of predicted outages. Our analysis will highlight dependencies of our analysis on storm severity, season and leaf condition.

Figure 6 summarizes the overall fit of QRF and BART aggregated by storm events. Both models show over-prediction for low impact events (<100 outages), while QRF also shows under-prediction for medium-high impact events (between 100 and 1000 outages). QRF (BART) exhibits coverage rates of prediction interval around 28% (36% to 60%), which were below our expectation (further analysis will reveal the cause of this phenomenon). We can see variations in performance across different leaf conditions and storm severities; both models did especially well at predicting hurricanes and for minor events with leaves off (winter).

First we will investigate the point estimate predictions of both models. Figure 7 (a), (b) and (c) illustrate how NSE varied for different spatial resolutions (grid, town, division) as a function of magnitude (outages per storm), while Figure 8 (a), (b) and (c) illustrate the same subject vs. different leaf conditions in boxplots. We see that the majority of storm events enjoy positive NSE values in Figure 7 and the 25% quantiles of NSE were close to or above 0 in Figure 8, indicating that the most of predictions were informative in predicting the spatial variability of outages. In addition, both models show that the NSE for the three resolutions generally exhibited a positive correlation between the accuracy of spatial variability modeling and the magnitude of the event outages (Figures 7). As expected (Section 3.3), we observed in both Figures 7 and 8 that for those events with positive NSE in grid cell resolution, the NSE's increased as the scale resolution became coarser (hence, we have many more events with NSE close to 1 at division resolution than at the grid resolution). Conversely, NSE tended to worsen at coarser scale aggregations for events where the models exhibited negative NSE at grid resolution. This "polarization" effect was so significant in division resolution that we should be cautious in using the aggregated results in this resolution, because the predicted outages may not reflect the true spatial distribution of outages for some storms. For all three resolutions, BART yielded better (more positive) NSE than QRF.

Figures 7 (d) and 8 (d) show how the relative error (RE) of aggregated storm total predictions varied in event magnitude and leaf condition. In the territory resolution (on the storm event scale), there is no NSE defined and instead we are more interested in how the point estimate performed in predicting the actual magnitude (outages per storm). The RE exhibited a negative correlation with magnitude in Figure 7, which suggests that both models were accurate for the most severe events. In Figure 8, BART frequently yielded RE's with smaller variance than QRF, which is consistent to its less spread-out predictions in Figure 6. Moreover, BART had the lowest (closest to zero) RE as well as highest NSE for the leaves off season (Figure 8). We attribute BART's improved winter (leaves off) season error metrics to the similarity of the data in the grouping (e.g. only minor events (no hurricanes), the ground is more likely to be frozen). In short, we conclude that BART yielded better point estimates than QRF because it had higher NSE and lower (more close to zero) RE than QRF.

4.2.2 Prediction Interval Results

We also examined the prediction intervals provided by BART and QRF with uncertainty ratio (UR), exceedance probability (EP) and rank histogram. Prediction intervals that are very wide offer no value to decision makers, because they suggest any amount of storm outage may occur; conversely, a too-narrow prediction interval may not give a useful estimate of the extent of possible outages. These widths of intervals are captured by UR in Figure 9 and a negative correlation between UR and magnitude was observed. Although narrow prediction intervals for high-magnitude events were favorable for their high certainty, the coverage rate becomes an important issue. In contrast to the UR, the actual value of outages exceeded the interval more frequently in severe events than in moderate events according to Figure 10 (a) and (b). We see a reverse trend of UR and EP vs. response magnitude. For BART, this is due to the homogeneity of variance assumption that made BART to offer intervals with similar absolute widths; for QRF, this is due to the fixed minimal number of instances in each terminal node that treated the nonzero or extreme responses the same as zero responses (more than 80% of responses as zeros at grid cell level in Figure 3 (a)). In practice, we look for prediction intervals that have acceptable small UR and stable EP which associated with the confidence level (e.g. a stable probability of 0.2 given 80% intervals in our case). This suggests that more flexible assumptions and settings are needed for BART and QRF to capture the variation in response magnitude.

During aggregation, UR reduced step by step from grid cell resolution to territory resolution (Figure 9), which is favorable. In contrast, the EP generally increased step by step via aggregation for both models (Figure 10), which led to the low interval coverage rates in Figure 6. In fact, QRF offered both low UR and low EP, implying superior performance in the highest resolution (i.e. grid cell). However, QRF's UR and EP became similar to or larger than BART's after aggregation, suggesting weakness in QRF's spatial aggregation. Since there are only four different divisions, EP can only be 0, 0.25, 0.5, 0.75 or 1 at division level in Figure 10 (c); Similarly, EP can only be 0 or 1 at storm event level in Figure 10 (d). Specifically, Figure 10 (c) and (d) address that QRF suffered more 1's of EP than BART.

To further elaborate the nature and issues of both models, we introduce rank histograms in Figures 11 and 12. In practice, a uniformly distributed rank histogram means the predicated distribution (including its quantiles and intervals generated by quantiles) reflects the variability of the actual response. Overall QRF (Figure 11 (a)) did well in grid cell resolution as evidenced by the near-uniform distributed rank histogram with a moderate under-prediction issue. In comparison, BART (Figure 12 (a)) produced biased predicted distribution by assuming normal distribution on Poisson-distributed actual outages (Figure 3 (a)) for grid cells. However, spatial aggregation appears to undermine the QRF prediction by accumulating biasedness (Figure 11). It is interesting to see BART (Figure 12) benefited a little from spatial aggregation. In fact, BART yielded better predictions for storm event totals, where the normal distribution becomes a better approximation of combined Poisson distributions (Figure 3). This explains why BART ended up with better interval coverage rates in Figure 6, even though QRF started from more accurate predicted distribution for grid cells. Note that biasedness also differs from location to location and aggregating locations with under-estimates and locations with over-estimates could result in complicated bias which is hard to predict. In conclusion, BART produced better prediction intervals for divisions and whole territory, while QRF did better for grid cells and towns.

4.3 Discussion

Similar to previous works in parametric modeling like Liu *et al.* [4], the models we utilized offer prediction intervals as well as point estimates of outages. Instead of simply identifying the potentials in quantifying prediction uncertainty, we took one more step in this study to evaluate QRF and BART with real-world data for their uncertainty measures. For hurricanes, BART model exceeded QRF in both predicting the outage magnitude (e.g. effective prediction intervals) and spatial variation of hurricane outages (Table II). BART under-predicted Irene by 1.9% and over-predicted Sandy by 2.4%, while the best ensemble decision tree model in our previous work (Wanik *et al.* [10]) over-predicted Irene by 11% and under-predicted Sandy by 4.4%. However, caution must be exercised when directly applying BART to storm outage modeling. Nateghi *et al.* [9] illustrated the weakness of BART when compared to random forest in survival analysis of hurricane outages where the response variable did not follow normal distribution. Most storms cause much less outages than hurricanes and even result in zero outage in some grid cells (Figure 3 (a)). Alternatively, QRF is promising in generating predictions and intervals without normality. Our analysis suggests that QRF suffered minor bias (Figure 11 (a)) in dealing with zero-inflated number of outages, while BART suffered significant bias (Figure 12 (a)). Compared to previous research (e.g. Guikema *et al.* [(8)]), we used real zero-inflated response variable based on storm events instead of simulated data or hurricanes, to suggest proper treatments to zeros. However, we found that BART was at least as good as QRF with respect to aggregated point estimates (Figure 7 and 8) and was better at generating aggregated prediction intervals (Figure 9 and 10). In short, different models could be utilized based on different interests or scales of application.

There are limitations for our study and results. Unlike Hurricane Ivan, studied by Nateghi *et al.* [(9)], Hurricanes Irene and Sandy did not make landfall in the service territory of our study. Since these storms were not at their peak when they impacted Connecticut, our research does not necessarily reflect the "worst case scenario" for outages. We did not include ice storms in this research due to their fundamentally different characteristics with other events in our database. The categorization of leaf conditions according to seasonal periods and spatial aggregation strategy according to geographic boundaries was based upon utility's demand to integrate the models with their emergency planning efforts.

5 CONCLUSIONS AND FUTURE WORK

This article has developed and validated outage prediction models for an electric distribution network. We incorporated high-resolution weather simulations, distribution infrastructure and land use for modeling storm outages using quantile regression forests (QRF) and Bayesian additive regression trees (BART). For hurricanes, BART model exceeded QRF in both predicting the outage magnitude and spatial variation of hurricanes. In our study, we found that outages caused by storms were not normally distributed and followed different distributions in different spatial resolutions. Hence, QRF was better at characterizing storm outages in high resolution, but did not aggregate well (from grid to division resolution). In contrast, BART did well at aggregating predictions for the storm outage total, but did not fully characterize the distribution of storm outages at higher resolution (e.g. grid resolution). In an operational context, utility companies might like to use maps of pre-storm outage predictions at the town resolution while also viewing a broader summary of the spatial variability, point estimates, and prediction intervals for the whole service territory. We suggest presenting the results from BART at coarser resolutions (e.g. division and service territory) and results from QRF for higher resolutions (e.g. grid and town) to best present the potential storm outages. Doing so will ensure that decision-makers get a complete idea of the overall severity of the event at a coarser resolution while also providing the detailed information supporting a pre-storm response at a higher resolution.

There are many opportunities for improvement in storm outage modeling on electric distribution networks. From a methodological point of view, both models could be modified to deal with the Poisson-distributed sparse (i.e. more than 80% as zeros) response variable. For QRF, empirical distributions for extreme observations are very different from the majority (mostly zeros). Varied treatments for the majority vs. real signals (i.e. outages) could help. By experimenting with the minimum number of observations required for the terminal node (according to overall magnitude of observations in the node), the accuracy of point estimates and prediction intervals may be improved. For BART, an application in generalized linear models (GLMs) with more flexible assumptions and link functions becomes very important, because Poisson-distributed data or zero-inflated data appear frequently in high-resolution analysis. For example, Poisson, negative binomial and zero-truncated normal with heterogeneous variance could perform as priors to assist BART in storm outage modeling. In addition to the two-staged model (similar to the classification-GAM model used by Guikema *et al.* [8]), a zero-inflated BART model optimized simultaneously for both zero-inflated classification and zero-truncated signals could be implemented. For spatial aggregation, the ideal unbiased prediction may not be available for every location, thus getting accurate predictions for another resolution based on biased results is challenging and important. There are already some advanced techniques (e.g. Reilly *et al.* [(41)]) to aggregate point estimates into multiple scales while eliminating bias and error by utilizing spatial patterns. Similar techniques to aggregate predicted distributions for each location could be investigated in a future study.

From a modeling point of view, the inconsistent performance of both models for varying season categories (tree-leaf condition) implies difficulties in predicting storm outages with leaves on trees. Future research may consider including additional effective explanatory variables that represent the localized tree conditions such as Leaf Area Index (LAI) [(42)], vegetation management (e.g. tree trimming) data or detailed tree density, location, height and species data to better capture this phenomenon.

REFERENCES

1. North American Electric Reliability Corporation (NERC). Events Analysis: System Disturbance Reports, 1992 - Continuing [Internet]. Atlanta (GA): North American Electric Reliability Corporation; 2010 [cited 2015 Jul 1]. Available from: `http://www.nerc.com/pa/rrm/ea/System%20Disturbance%20Reports%20DL/Forms/AllItems.aspx`.

2. Liu H, Davidson RA, Rosowsky DV, Stedinger JR. Negative binomial regression of electric power outages in hurricanes. Journal of Infrastructure Systems, 2005; 11(4):258-267.

3. Guikema SD, Davidson RA, Liu H. Statistical models of the effects of tree trimming on power system outages. Power Systems, IEEE Transactions on, 2006; 11(3):1549-1557.

4. Liu H, Davidson RA, Apanasovich TV. Spatial generalized linear mixed models of electric power outages due to hurricanes and ice storms. Reliability Engineering & System Safety, 2008; 93(6):897-912.

5. Han SR, Guikema SD, Quiring SM, Lee KH, Rosowsky D, Davidson RA. Estimating the spatial distribution of power outages during hurricanes in the Gulf coast region. Reliability Engineering & System Safety, 2009; 94(2):199-210.

6. Guikema SD, Quiring SM, Han SR. Prestorm estimation of hurricane damage to electric power distribution systems. Risk analysis, 2010; 30(12):1744-1752.

7. Nateghi R, Guikema SD, Quiring SM. Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes. Risk analysis, 2011; 31(12):1897-1906.

8. Guikema SD, Quiring SM. Hybrid data mining-regression for infrastructure risk assessment based on zero-inflated data. Reliability Engineering & System Safety, 2012; 99:178-182.

9. Nateghi R, Guikema SD, Quiring SM. Forecasting hurricane-induced power outage durations. Natural Hazards, 2014; 74(3):1795-1811.

10. Wanik DW, Anagnostou EN, Hartman BM, Frediani MEB, Astitha M. Storm outage modeling for an electric distribution network in Northeastern US. Natural Hazards, 2015.

11. Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda MG, Huang XY, Wang W, Powers JG. A description of the advanced research WRF version 3 [Internet]. NCAR Technical Note, 2008 [cited 2015 Feb 15]; NCAR/TN-475+STR. Available from: `http://www2.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf`.

12. National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce. NCEP global forecast system (GFS) analyses and forecasts [Internet]. Boulder (CO): Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory; 2007 [cited 2015 Feb 15]. Available from: `http://rda.ucar.edu/datasets/ds084.6/`.

13. National Centers for Environmental Prediction, National Weather Service, NOAA. NCEP ADP Global Upper Air and Surface Weather Observations (PREPBUFR format), May 1997 - Continuing [Internet]. Boulder (CO): Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory; 2008 [cited 2015 Feb 15]. Available from: `http://rda.ucar.edu/datasets/ds337.0/`.

14. Sellier DD. Crown structure and wood properties: Influence on tree sway and response to high winds. American Journal of Botany, 2009; 96(5):885-896.

15. Ciftci C, Arwade SR, Kane B, Brena SF. Analysis of the probability of failure for open-grown trees during wind storms. Probabilistic Engineering Mechanics, 2014; 37:41-50.

16. Quiring SM, Zhu L, Guikema SD. Importance of soil and elevation characteristics for modeling hurricane-induced power outages. Natural hazards, 2011; 58(1):365-390.

17. Center for Land Use Education and Research, College of Agriculture and Natural Resources, University of Connecticut. Connecticut's Changing Landscape (CCL) Basic Land Cover Information, 1990 - Continuing [Internet]. Haddam (CT): Center for Land Use Education and Research; 2006 [cited 2015 Feb 15]. Available from: http://clear.uconn.edu/projects/landscape/.

18. Breiman L. Random forests. Machine learning, 2001; 45(1):5-32.

19. Meinshausen N. Quantile regression forests. The Journal of Machine Learning Research, 2006; 7:983-999.

20. Meinshausen N. quantregForest: quantile regression forests. R package version 0.2-3 [Internet]. Vienna (Austria): CRAN, Institute for Statistics and Mathematics, Vienna University of Economics and Business; 2012 [cited 2015 Feb 15]. Available from: http://cran.r-project.org/web/packages/quantregForest/.

21. Liaw A, Wiener M. The randomForest package. R News, 2002; 2(3):18-22.

22. Juban J, Fugon L, Kariniotakis G *et al.* Probabilistic short-term wind power forecasting based on kernel density estimators. European Wind Energy Conference and Exhibition, EWEC; 2007 May 7-10; Milan, Italy.

23. Francke T, López-Tarazón JA, Vericat D, Bronstert A, Batalla RJ. Flood-based analysis of high-magnitude sediment transport using a non-parametric method. Earth Surface Processes and Landforms, 2008; 33(13):2064-2077.

24. Zimmermann A, Francke T, Elsenbeer H. Forests and erosion: Insights from a study of suspended-sediment dynamics in an overland flow-prone rainforest catchment. Journal of Hydrology, 2012; 428:170-181.

25. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Technical Report, Department of Mathmatics and Statistics, Acadia University, 2005 July.

26. Chipman HA, George EI, McCulloch RE *et al.* BART: Bayesian additive regression trees. The Annals of Applied Statistics, 2010; 4(1):266-298.

27. Hastie T, Tibshirani R *et al.* Bayesian backfitting (with comments and a rejoinder by the authors). Statistical Science, 2000; 15(3):196-223.

28. Chipman HA, George EI, McCulloch RE. Bayesian CART model search. Journal of the American Statistical Association, 1998; 93(443):935-948.

29. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 2001; 1189-1232.

30. Chipman HA, McCulloch RE. BayesTree: Bayesian methods for tree based models. R package version 0.3-1 [Internet]. Vienna (Austria): CRAN, Institute for Statistics and Mathematics, Vienna University of Economics and Business; 2009 [cited 2015 Feb 15]. Available from: `http://cran.r-project.org/web/packages/BayesTree/`.

31. Pratola MT, Chipman HA, Gattiker JR, Higdon DM, McCulloch RE, Rust WN. Parallel Bayesian additive regression trees. Journal of Computational and Graphical Statistics, 2014; 23(3):830-852.

32. Kapelner A, Bleich J. bartMachine: Machine Learning with Bayesian additive regression trees. R package version 1.1.1 [Internet]. Vienna (Austria): CRAN, Institute for Statistics and Mathematics, Vienna University of Economics and Business; 2014 [cited 2015 Feb 15]. Available from: `http://cran.r-project.org/web/packages/bartMachine/`.

33. Blattenberger G, Fowles R. Avalanche forecasting: using Bayesian additive regression trees (BART). Demand for communications services–insights and perspectives. New York: Springer; 2014. Part III, Empirical Applications: Other Areas; p. 211-227.

34. Gehlke CE, Biehl K. Certain effects of grouping upon the size of the correlation coefficient in census tract material. Journal of the American Statistical Association, 1934;

29(185A):169-170.

35. Openshaw S, Taylor PJ. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. Statistical Applications in the Spatial Sciences, 1979; 21:127-144.

36. Nash JE, Sutcliffe JV. River flow forecasting through conceptual models part IA discussion of principles. Journal of Hydrology, 1970; 10(3):282-290.

37. Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 2007; 50(3):885-900.

38. Özkaynak H, Frey HC, Burke J, Pinder RW. Analysis of coupled model uncertainties in source-to-dose modeling of human exposures to ambient air pollution: A PM 2.5 case study. Atmospheric Environment, 2009; 43(9):1641-1649.

39. Hamill TM. Interpretation of rank histograms for verifying ensemble forecasts. Monthly Weather Review, 2001; 129(3):550-560.

40. Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. The Annals of Statistics, 1981:1196-1217.

41. Reilly A, Guikema SD. Bayesian Multiscale Modeling of Spatial Infrastructure Performance Predictions with an Application to Electric Power Outage Forecasting. Journal of Infrastructure Systems, 2014; 21(2):04014036.

42. NASA Earth Observations (NEO). Leaf Area Index (LAI), 2000 - Continuing [Internet]. Greenbelt (MD): EOS Project Science Office, NASA Goddard Space Flight Center; 2015 [cited 2015 Sep 15]. Available from: `http://neo.sci.gsfc.nasa.gov/view.php?datasetId=MOD15A2_M_LAI`.

Table I: Explanatory Variables Included in Modeling

Variable	Description	Type	Unit
Wind10m	Sustained wind speed at 10 meters	Numerical	m/s
Gust	Wind speed of gust at 10 meters	Numerical	m/s
WStress	Wind stress	Numerical	-
wgt9	Duration of 10m wind greater than 9m/s	Numerical	hour
ggt18	Duration of gust greater than 18m/s	Numerical	hour
ggt27	Duration of gust greater than 27m/s	Numerical	hour
PreRate	Precipitation rate	Numerical	mm/hr
TotPrec	Total accumulated precipitation	Numerical	mm
Temp	Temperature	Numerical	°C
SoilMst	Soil moisture	Numerical	kg/kg
SnoWtEq	Snow water equivalent (only for winter)	Numerical	kg/m^2
sumAssets	Sum of assets (infrastructure)	Numerical	count
PercDeveloped	Percentage of urban area	Numerical	%
PercConif	Percentage of coniferous trees	Numerical	%
PercDecid	Percentage of deciduous tress	Numerical	%
seasoncat	Leaves on (summer: Jun to Sep); Leaves off (winter: Dec to Mar); Transition (Oct, Nov, Apr and May).	Categorical	-

Table II: Comparison of QRF and BART with Hurricane Validation Data

Hurricane (Outages)	**Model**	**Predicted**	**Pred. Interval**	**MAE (By Town)**	**RMSE (By Town)**
Irene (4890)	QRF	4542	(4311, 4666)	8.86	15.70
	BART	4795	(4688, 4898)	6.12	9.43
	MEAN	5200	-	23.95	35.25
Sandy (5052)	QRF	5060	(4674, 5110)	6.91	11.02
	BART	5171	(5039, 5302)	5.50	8.02
	MEAN	5121	-	28.89	52.03

Figure 1: Spatial Resolutions: 2-km Grid Cell, Town, Division and Territory. (Grid cells without infrastructure or outside the service territory are excluded.)

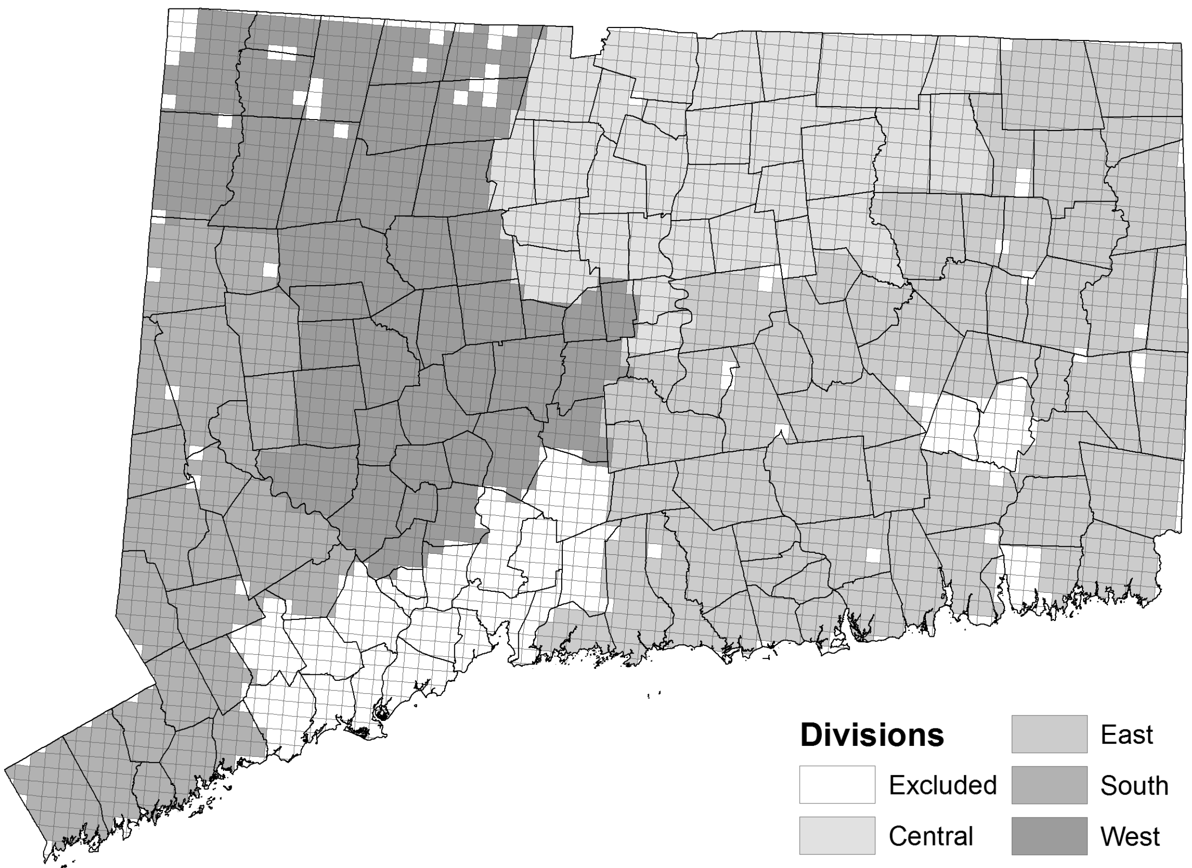

Figure 2: Storm Events Simulation: Weather Research and Forecasting Model Nested Domains in 18 km, 6 km and 2 km grids.

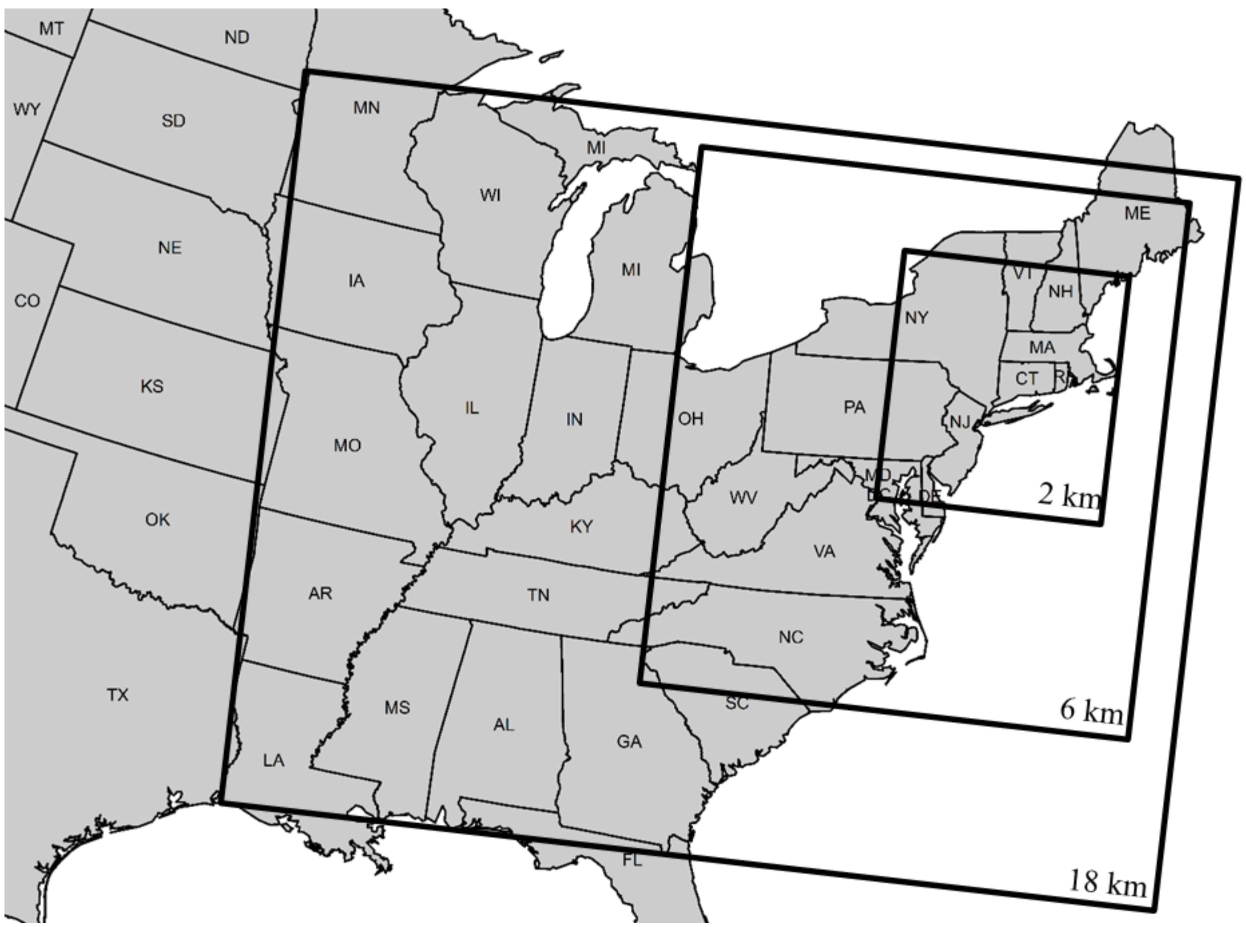

Figure 3: Density of Outages in Different Resolutions: (a) Grid Cell, (b) Town, (c) Division, (d) Territory. (Dotted lines stand for gaussian kernel density estimations.)

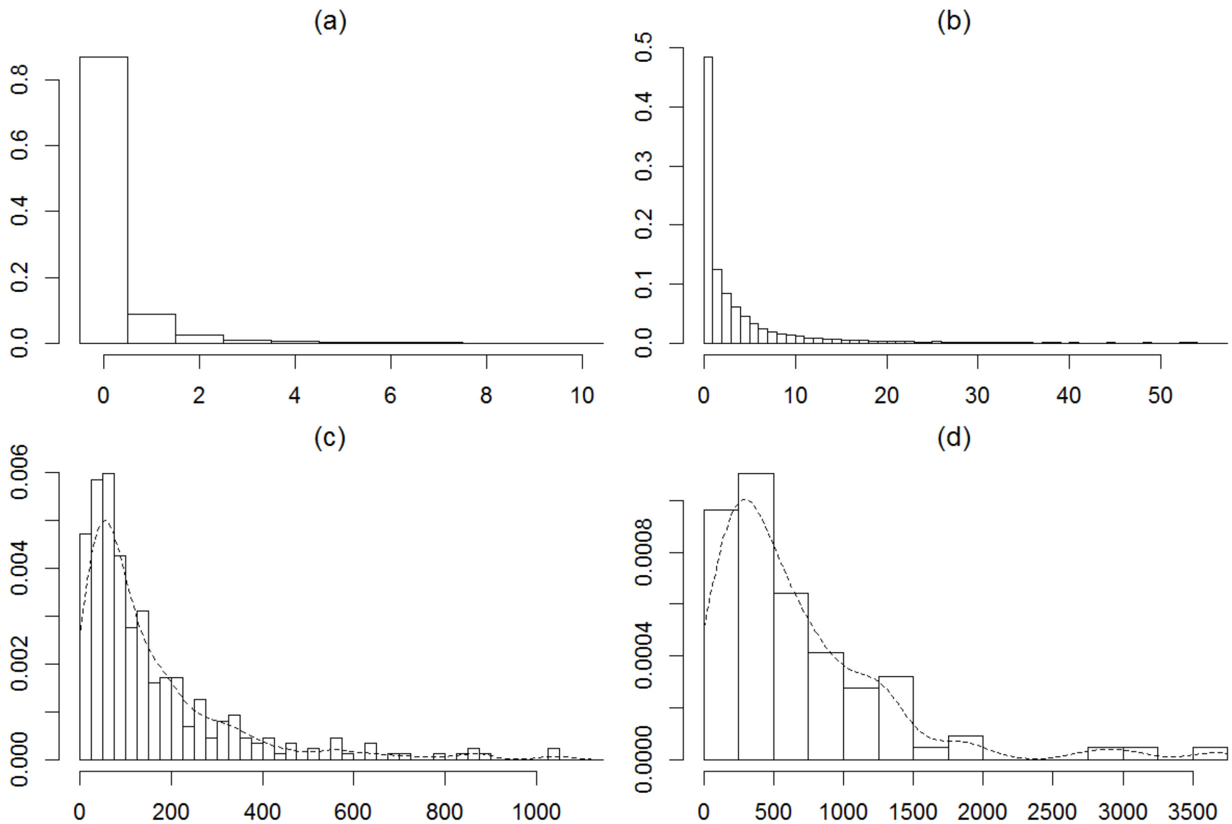

Figure 4: Comparison of QRF and BART in Town Resolution: (a) Irene, (b) Sandy. (80% prediction intervals are given, as well as their coverage rates.)

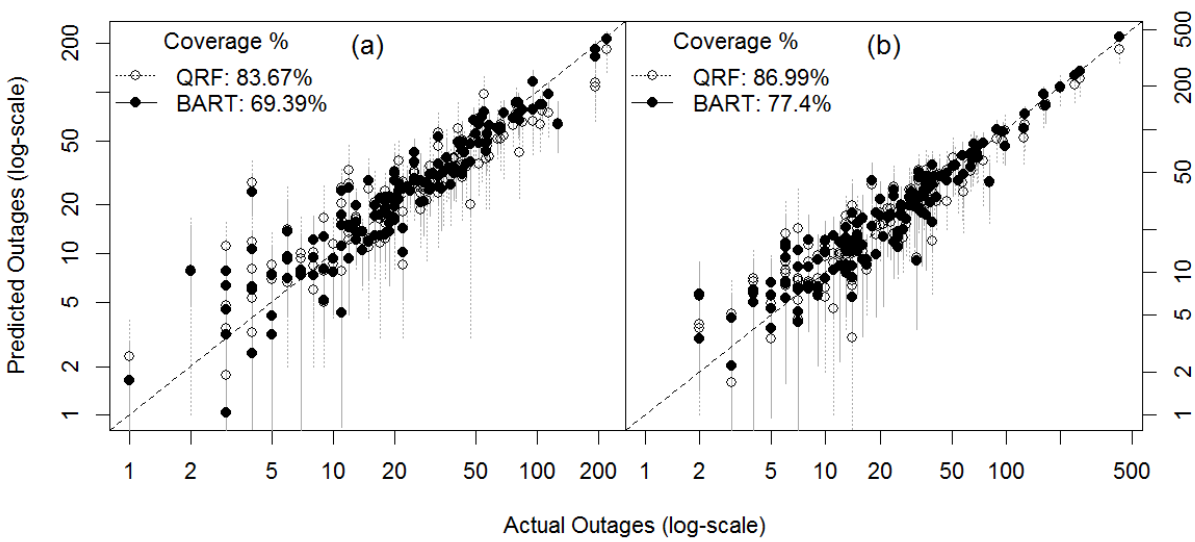

Figure 5: Comparison of QRF and BART in Modeling Spatial Variability: Total Number of Outages by Town for (a) Actual Number of Irene, (b) Actual Number of Sandy, (c) QRF Validation of Irene, (d) QRF Validation of Sandy, (e) BART Validation of Irene and (f) BART Validation of Sandy.

Figure 6: Predictions for Each Storm in Validation Dataset. (80% prediction intervals are given, as well as their coverage rates.)

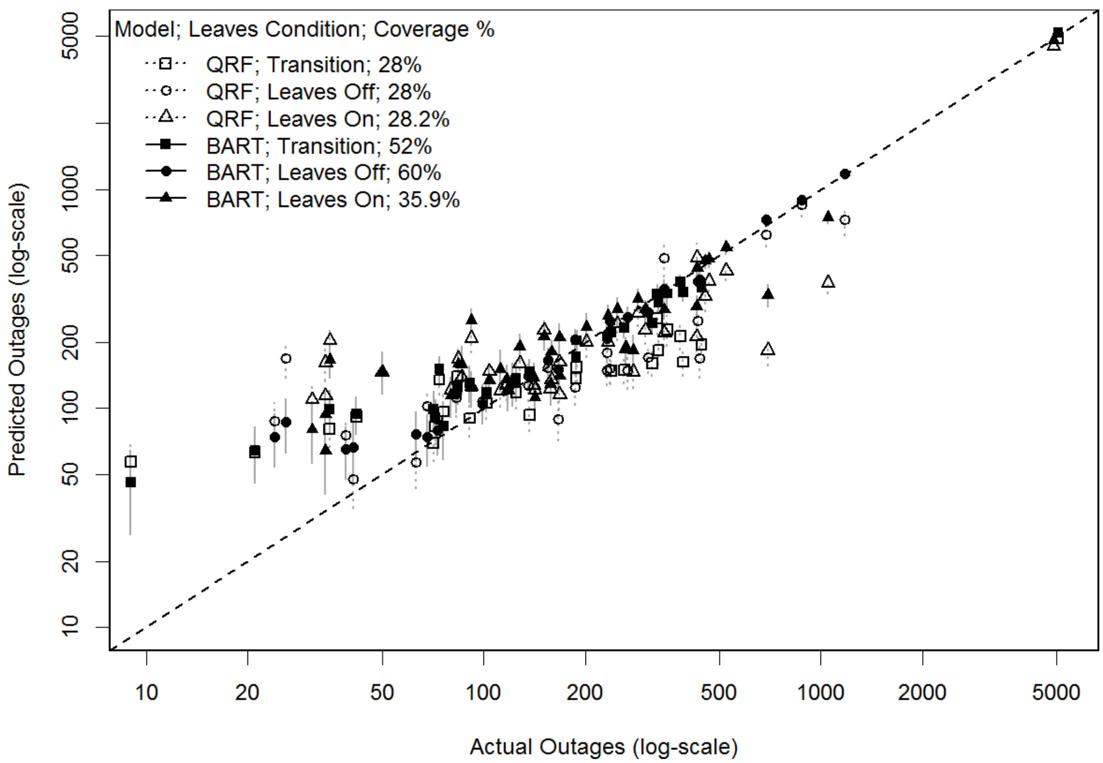

Figure 7: Nash-Sutcliffe Efficiency (NSE) or Relative Error (RE) for Each Storm in Different Resolutions: (a) NSE for Grid Cells, (b) NSE for Towns, (c) NSE for Divisions, (d) RE for Territory.

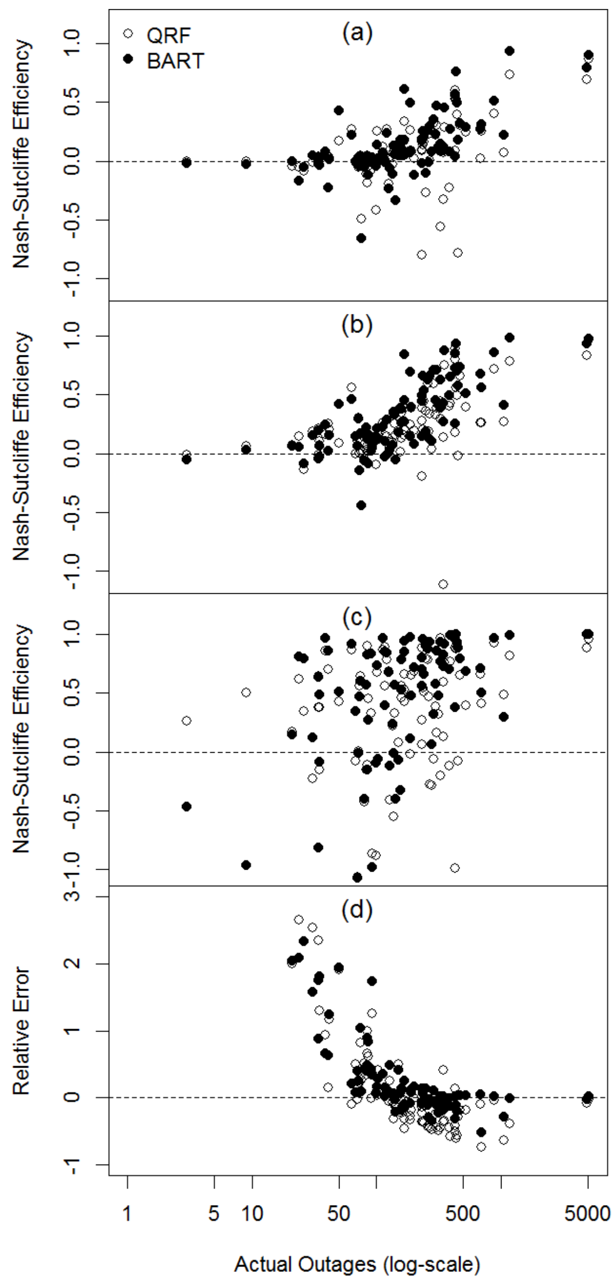

Figure 8: Nash-Sutcliffe Efficiency (NSE) or Relative Error (RE) for Each Season in Different Resolutions: (a) NSE for Grid Cells, (b) NSE for Towns, (c) NSE for Divisions, (d) RE for Territory.

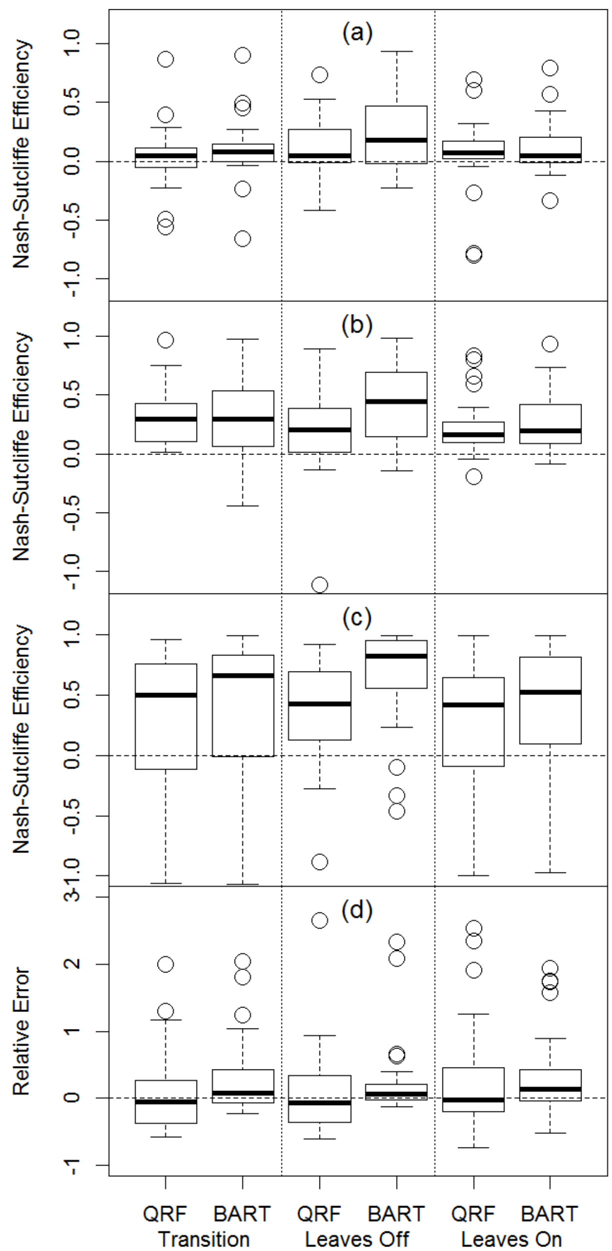

Figure 9: Uncertainty Ratio (UR) for Each Storm in Different Resolutions: (a) Grid Cell, (b) Town, (c) Division, (d) Territory.

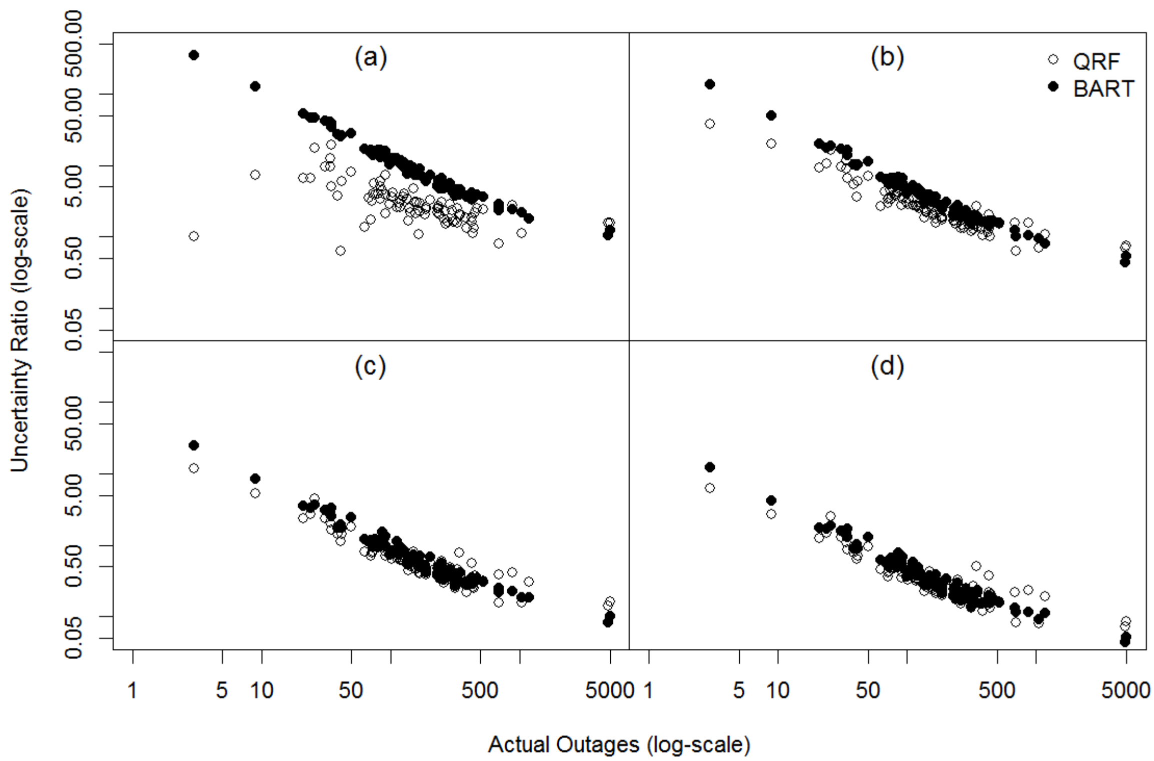

Figure 10: Exceedance Probability (EP) for Each Storm in Different Resolutions: (a) Grid Cell, (b) Town, (c) Division, (d) Territory. (A small amount of noise is add in (c) and (d) to avoid overlap of points.)

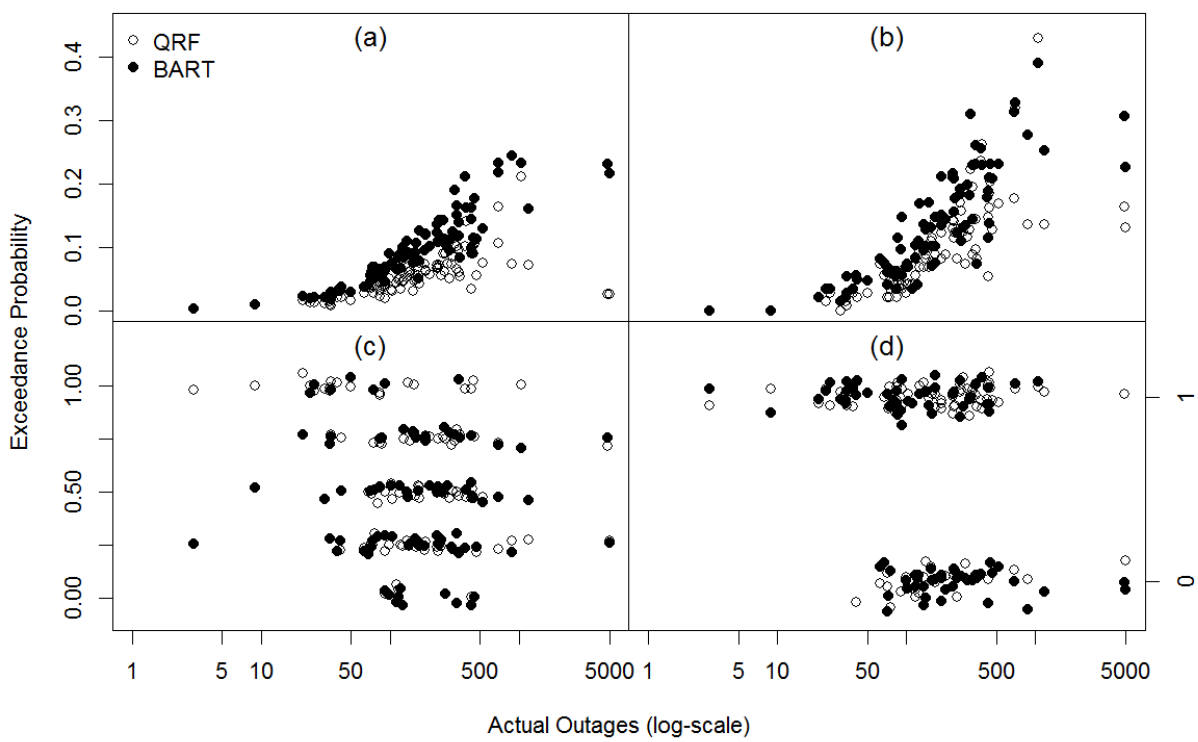

Figure 11: Rank Histogram of QRF Predictions in Different Resolutions: (a) Grid Cell, (b) Town, (c) Division, (d) Territory.

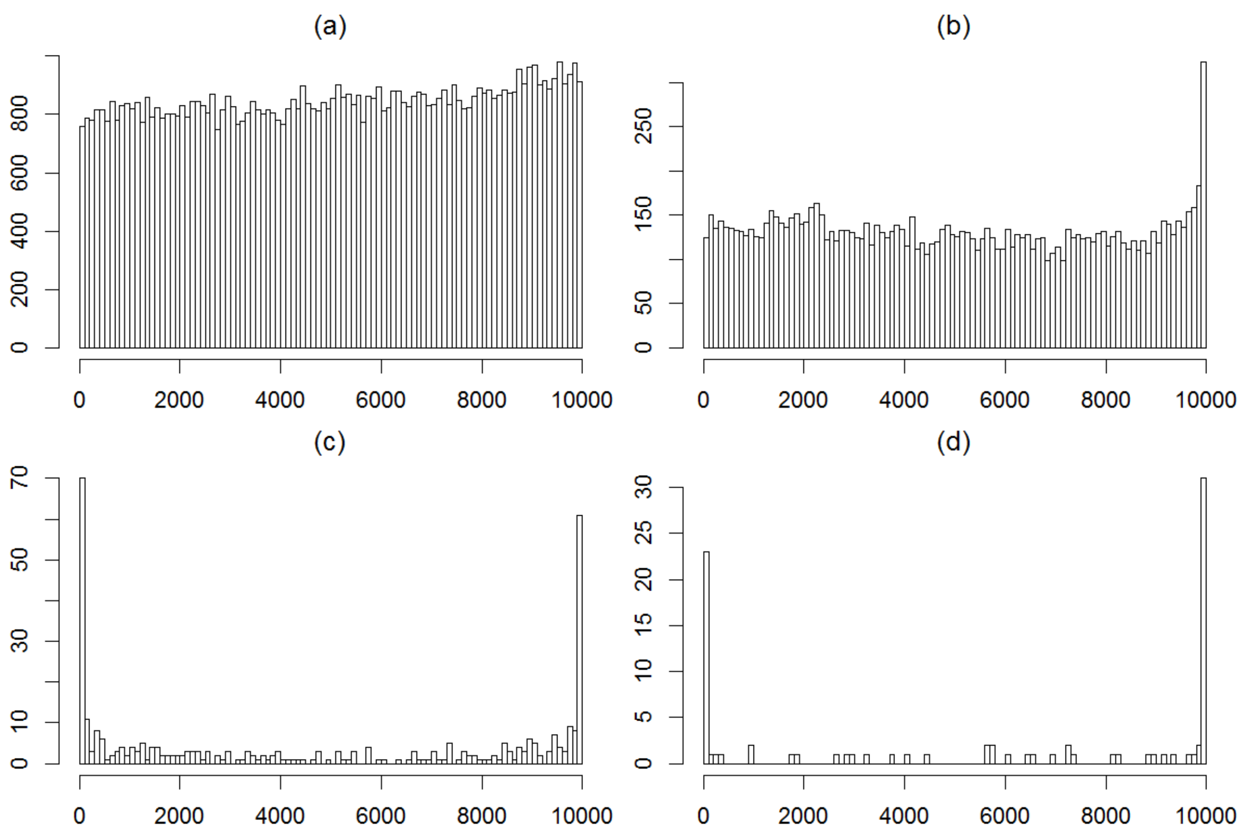

Figure 12: Rank Histogram of BART Predictions in Different Resolutions: (a) Grid Cell, (b) Town, (c) Division, (d) Territory.