Journal of Renewable and Sustainable Energy

Bayesian and non-Bayesian regression analysis applied on wind speed data

Cite as: J. Renewable Sustainable Energy **13**, 053303 (2021); https://doi.org/10.1063/5.0056237
Submitted: 07 May 2021 . Accepted: 13 September 2021 . Published Online: 11 October 2021

Vincent Tanoe, Saul Henderson, Amir Shahirinia, et al.

Journal of Renewable and Sustainable Energy ARTICLE scitation.org/journal/rse

Bayesian and non-Bayesian regression analysis applied on wind speed data

Cite as: J. Renewable Sustainable Energy **13**, 053303 (2021); doi: 10.1063/5.0056237
Submitted: 7 May 2021 · Accepted: 13 September 2021 ·
Published Online: 11 October 2021

Vincent Tanoe,[1,a)] Saul Henderson,[2] Amir Shahirinia,[3] and Mohammad Tavakoli Bina[4]

AFFILIATIONS

[1]Computer Science and Engineering Department, University of the District of Columbia, 4200 Connecticut Avenue NW, Washington, DC 20008, USA
[2]Electrical and Computer Engineering Department, University of the District of Columbia, 4200 Connecticut Avenue NW, Washington, DC 20008, USA
[3]Electrical and Engineering Department, University of the District of Columbia, 4200 Connecticut Avenue NW, Washington, DC 20008, USA
[4]K. N. Toosi University of Technology, Tehran, Iran

[a)]**Author to whom correspondence should be addressed:** vincent.tanoe@udc.edu

ABSTRACT

Statistical methods are widely used to analyze the relationship between several independent variables (predictors) and a dependent variable. As wind energy rapidly becomes an important source of renewable energy, it is prudent to deeply evaluate any potential existing relationships among the data. This paper aims to apply the frequentist statistical approach, namely, non-Bayesian and the Bayesian approach, to multiple linear regression to wind speed data to investigate the differences between the two methodologies. This study uses the NREL wind speed data from fifteen different wind farms. In the proposed study, a correlation matrix was implemented to select the significantly correlated variables among all and use it as the dependent variable. This method is followed by a Random Forest machine learning technique for feature selection and considering the most important features that will be used for the Bayesian and non-Bayesian regression models. We first run a multiple linear regression (non-Bayesian regression model) in which we apply the variance inflation factor to detect any multicollinearity problem to get the fitted model. We then apply the Bayesian approach to the fitted model to analyze the relationship between the dependent and independent variables. The results from both non-Bayesian and the Bayesian approaches show close coefficients and parameters estimations. Moreover, using different wind speed data sample sizes of hourly, daily, and weekly data, we found that the daily data provide a strong coefficient estimator and the highest R-squared compared to the hourly and weekly data.

NOMENCLATURE

Abbreviations

ECDF	empirical cumulative distribution function
K-S	Kolmogorov–Smirnov
MCMC	Monte Carlo Markov chain
NREL	National Renewable Energy Laboratory
RF	Random Forest
VIF	variance inflator factor

Parameters

D_n	K–S test statistics
E_N	ECDF function
F	the theoretical cumulative distribution
fi_i	importance of features
H_t	represents Humboldt data at time t
$\widehat{H_t}$	the predicted value of H_t
$L(\beta_o, \beta_i, \phi)$	likelihood functions
$P(A)$, $P(B$	marginal probability
ϕ	gamma prior

I. INTRODUCTION

Over the past 10 years, the wind has been used as a domestic source of energy. Wind farms are very important as they provide wind energy. Wind power shows a significant capacity growth of 15% in the United States every year according to the Office of Energy Efficiency and Renewable Energy. This fastest-growing energy source is

explained by many factors, such as effective cost, jobs creations, competitiveness among industries, and sustainability. Wind energy is generated via wind turbines and wind farms. The benefit of wind turbines can be seen in the reduction of the amount of electricity generated from fossil fuels, which in turn lowers air pollution and carbon dioxide emissions. This significant increase in wind farm usage has captivated research attention. Scientists and engineers analyze wind speed and wind turbines to accurately measure the power output. In Ref. 1, the impact of wind speed trends and 30-year variability about hydroelectric reservoir inflows on wind power in the Pacific Northwest was analyzed. The authors used British Columbia and the Pacific Northwest as a case study and found that clean energy and self-sufficiency policies in British make the benefits of increased generation during low streamflow periods particularly large. Other research emphasized by scientists involve power forecasting,[2] wind turbine detection,[3,4] and stochastic economic dispatch.[5,6] In Ref. 7, to capture the uncertainty effects in the technical decisions of optimal scheduling, a stochastic approach based on unscented transform was developed to handle the forecast error in electrical and thermal energy demands, market energy prices related to the different energy layers, and the output power forecast error in the renewable energy sources. The authors applied a novel reinforcement learning-based approach to find a near-optimal solution and to facilitate the searching process with a trivial computational burden. For the investigation of the optimal management of multicarrier water and energy system, Zot *et al.*[8] used reinforcement learning and unscented transform. Their reinforcement learning-based approach was devised for finding a near-optimal solution and facilitates the searching process with a trivial computational burden. Their simulation results indicated that the proposed cooperation approach minimized both the operation and the investment cost substantially with an efficient computational burden based on the advanced features coming out of the proposed reinforcement learning approach. Other authors used a stochastic machine learning-based approach for observability enhancement of automated smart grids.[9] In their proposed stochastic approach, the authors presented a strategy occurring in several stages to micro-synchrophasor unit positioning based on the load level and demand in the system and based on the predetermined sectionalizing and tie switches. In Ref. 10, an effective stochastic framework for smart coordinated operation of wind park and energy storage unit has been analyzed. The authors proposed a stochastic transmission switching integrated interval robust chance-constrained approach to assess the operation of a wind park-energy storage system in a day-ahead electricity market considering the system's technical constraints.

Analyzing wind speed data is quite challenging due to their nature of uncertainty. Uncertainty regarding wind speed data has been evaluated through probability distributions. The uncertainty of the wind speed primarily views the distributions of the wind speed over a wind farm as being homogeneous. However, the uncertainty about these wind speed models has not yet been considered. In this study, we propose a method of analysis, where we conduct the step-by-step implementation of both Bayesian and non-Bayesian regression models analysis applied on wind speed data.

Although many models have been applied to analyze the dependencies between the wind speed, a holistic important aspect of evaluating the significant impact among wind speed data should be addressed[2–6] This regression method approach has not been considered yet. However, to attract more attention effectively on the significant impact among the wind speed data, one can ask the following question:

If a new development is designed to analyze independent variables, what can we expect in the dependent variable as a result? Such a question requires a statistical concept of Bayesian and non-Bayesian regression modeling. The Bayesian model is a statistical model based on Bayes theorem. It is a model where probabilities are used to represent all uncertainty within the model, and both uncertainty regarding the output (dependent variable) and the uncertainty of the input (independent variables) to the model.[11] On the other side, the non-Bayesian regression model is a linear regression model that analyzes the relationship between an output (dependent variable) and the input variables (independent variables). The application of Bayesian methods in wind speed data is essential for improving the accuracy and reliability of wind resource estimation and short-term forecasts. In the case of the non-Bayesian model, it will help us determine the most important explanatory variables and how they can be used for predicting or forecasting the output.

The remainder of the paper is organized as follows: Sec. II provides the proposed method and an overview of the theoretical and empirical research on wind speed data. Section III gives the literature review of the related existing studies. In Sec. IV, the data and methodology are covered. Section V presents how the experimental results of the statistics and machine learning models are estimated from the data and the results are interpreted. Section VI concludes the paper and future works.

II. THE PROPOSED METHOD

In this paper, the wind speed model analyses will be performed in the following steps:

- Determine the most correlated variable from all wind speed data variables (15) using the correlation matrix and rank from the most to the lowest correlated variables
- Use the top-ranked correlated variable as the dependent variable
- Apply the random forest method on the rest of the variables to select the features (explanatory variables) and use the first selected feature as the most important independent variable.
- Apply the variance inflation factor (VIF) on the regression for multicollinearity and select the fitted model
- Run the non-Bayesian and Bayesian and approaches using the fitted model to predict and compare the results.

To better analyze the uncertainty about the wind speed data and provide a better predictive model, we split the data into three different categories of the dataset, namely, large, medium, and small. This paper uses 15 wind speed data variables from different sites across the US from the National Renewable Energy Laboratory (NREL) website.[12] The variables are in hourly data (large dataset, 8760 observations), daily (medium dataset, 365 observations), and weekly (small dataset, 53 observations).

III. LITERATURE REVIEW

Analyzing the relationship between several independent variables (predictors) and a dependent variable is widely used by the researchers, scholars, and decision-makers to predict or forecast future events. Multiple regression is designed to investigate the relative influences of the independent variables on dependent variables. To implement this technique, one need is to define a hypothesis test. This type of analysis is classified as frequentist statistics and is considered a non-Bayesian

approach. Unlike the non-Bayesian approach, the Bayesian approach uses probabilities and prior distributions, likelihood, and predictive posteriors. In the Bayesian approach, data are observed and prior believes are updated to form a posterior distribution. This approach is based on Bayes theorem, which is used to calculate the conditional probabilities. The Bayes theorem is mathematically described as the follows:[19]

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1}$$

where A and B are events, and $P(B)$ is not equal to zero. $P(B|A)$ is the conditional probability where the likelihood of event A is occurring given that B is true. $P(B|A)$ is also a conditional probability where the likelihood of event B is occurring given that A is true. Finally, $P(A)$ and $P(B)$ are the probabilities of observing A and B, respectively, and are also known as the marginal probabilities. A and B must be different.

The difference between the two methods is that the non-Bayesian or frequentist statistic does not explicitly involve a prior while the Bayesian does. Several research studies have used the two methods for different purposes. In Ref. 13, the authors used the Bayesian learning model approach to model the energy performance of residential buildings. The authors compared the Bayesian approach and the ordinary least squared (OLS) method which is also considered as a frequentist statistic and found that the Bayesian method outperformed the OLS based on certain criteria, including root mean square error (RMSE), mean-absolute-percentage error (MAPE), and median absolute deviation (MAD). Many statisticians in the past found that the Bayesian technique was not a satisfactory method to use because the choice of prior distribution was unfounded and varied among statisticians. Today, the advancement of computers allows for the implementation of the Bayesian model approach. Bayesian modeling is widely used as it allows for the interpretation of probability as a measure of the degree of belief concerning actual data observed.[14] To analyze the change point models in hydrometeorological variables, researchers in Ref. 11 applied Bayesian multivariable linear regression to allow simultaneous single change point detection in a multivariate sample and accounted for missing data in the response variables and/or in the explicative variables. Moreover, wind speed data forecasting can serve a wide spectrum of purposes, including scheduling of a power system and dynamic control of structures. In Ref. 15, the probabilistic forecast of wind speed based on a Bayesian emulator using monitoring data has been analyzed. As the result of their studies, they found that the Bayesian emulator approach not only maintained the data-driven property which guarantees its high flexibility in modeling the complexity of the target system but also allows for efficiency. Probabilistic evaluation of the wind speed in terms of the predictive mean and variance. In this research, we have emphasized using frequentist statistics (linear regression) and the Bayesian method to study the wind speed data and the uncertainties associated with the developed models.

IV. METHODOLOGY AND DATA

This paper uses 15 wind speed data from different sites across the US. As stated, the variables are hourly (8760), daily (365), and weekly (52) formats. Daily and weekly data are obtained from the hourly data by taking the average over the days and weeks. The descriptive statistics of the hourly data are shown in Table I.

A. Correlation matrix and features selections

1. Correlation matrix

To begin our analysis, we have preprocessed all 15 variables by using the correlation matrix. The matrix is used in this study to select the correlation coefficients among these 15 sets of variables. The correlation measures the degree of associated linearity between two continuous variables. A positive correlation shows that if the value of a continuous variable increases the value of the other continuous variable increases, too. In contrast, a negative correlation indicates that as the value of one of the continuous variables increases, the value of the other continuous variable decreases. The standard method of correlation (Pearson) is used to evaluate the correlation. The Pearson's formula is described as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2}$$

where r is the Pearson's correlation coefficient that can only take values between -1 and 1, while x and y represent the two continuous variables.

We imposed a cutoff of 0.5 in our correlation matrix to obtain the rank of correlation among the data. The goal of this method is to determine which variable could be considered as the dependent variable for the regression model. From the correlation matrix method, 9 out of the 15 variables were ranked with a correlation value of at least 0.5. Figure 1 shows the alignment of the data based on their associativity linear relationship with the other data. The Humboldt wind site is the first variable that we used as the dependent variable.

After the correlation matrix was applied and the highest dependent (Humboldt) variable was determined, we continued our data preprocessing by using the Random Forest machine learning technique to select the features based on their important effect on the selected dependent variable.

2. Features selections

Feature selection, or importance, is a statistical technique that is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability is then calculated by the number of samples that reach the node and then divided by the total number of samples. The higher the value, the more important the feature.[16]

The RF technique was used for the feature selection process. RF is an ensemble technique that constructs many individual decision trees as training. The final predictions are obtained from all trees, and then, the mode of the classes for classification or the mean prediction regression is acquired.[17] For each decision tree, we first assume only two binary trees:[17]

$$ni_j = w_j I_j - w_{left(j)} I_{left(j)} - w_{right(j)} I_{right(j)}, \tag{3}$$

where ni_j represents the importance of node j, w_j is the weighted number of samples that reach node j, I_j is the impurity value of the node j, $left(j)$ represents the child node from the left split on node j, and finally $right(j)$ represents the child node from right split on node j.

We then calculate the importance of each feature on the decision tree using the following:[17]

TABLE I. Descriptive statistics of hourly wind speed data.

Summary	Bear Creek	Frey Farm	Criterion Wind Park	Ned Power	Humboldt
Count	8760	8760	8760	8760	8760
Mean	7.66	5.73	8.23	7.66	7.66
Std	3.64	3.32	4.42	3.94	3.58
Min	0.21	0.04	0.15	0.15	0.12
25%	5.03	3.19	4.85	4.76	5.04
50%	7.27	5.35	7.6	7.15	7.36
75%	9.86	7.95	10.98	10.09	9.89
Max	28.51	22.17	26.02	23.28	26.66
Summary	Locust Ridge	Roth Rock	Talbot	Mountaineer	Buffalo Mountain
Count	8760	8760	8760	8760	8760
Mean	5.41	7.05	6.63	8.75	7.55
Std	3.36	3.81	3.25	5.42	4.02
Min	0.05	0.1	0.1	0.16	0.16
25%	3.03	4.19	4.19	4.5	4.47
50%	4.85	6.58	6.4	7.68	7
75%	7.15	9.41	8.81	11.86	10.06
Max	26.23	22.3	23.95	29.1	24.99
Summary	Bit Works	Mt Peak Utility	Anacacho	Dry Lake	Kingman
Count	8760	8760	8760	8760	8760
Mean	7.04	7.73	7.37	6	7.16
Std	3.48	3.48	3.18	3.43	3.7
Min	0.08	0.1	0.06	0.09	0.06
25%	4.46	4.98	4.97	3.31	4.36
50%	6.92	7.85	7.32	5.52	6.89
75%	9.43	10.36	9.71	8.23	9.61
Max	20.66	19.15	21.61	25.12	21.19

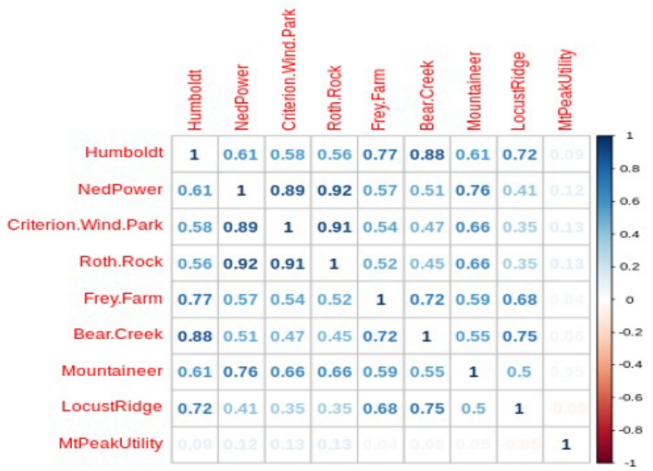

FIG. 1. Correlation matrix of wind speed data.

$$fi_i = \frac{\sum_{j:\text{node } j \text{ splits on feature } i} ni_j}{\sum_{k \in \text{all nodes}} ni_k}, \tag{4}$$

where fi_i is described as the importance of features.

To normalize the value between 0 and 1, we divide the sum of all feature importance values from the following:[17]

$$normfi_i = \frac{fi_i}{\sum_{j \in \text{all features}} fi_j}. \tag{5}$$

Finally, at the RF level, the feature importance is determined by the average of overall trees. The sum of the feature's importance value of each tree is calculated then divided by the total number of trees applied by this formula,[17]

$$RFfi_i = \frac{\sum_{j \in \text{all trees}} normfi_{ij}}{T}, \tag{6}$$

where $RFfi_i$ is the importance of feature i calculated from all trees in the RF model and T the total number of trees.

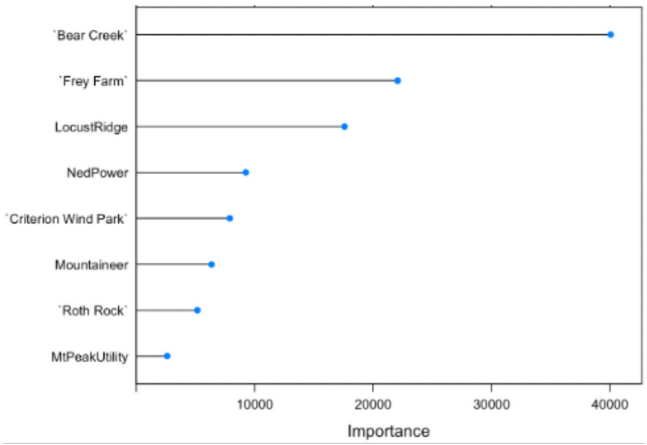

FIG. 2. Feature's selections of wind speed data.

The features based on their importance regarding the dependent variables are shown in Fig. 2. From the highest ranked to the lowest-ranked, we have Bear Creek, Frey Farm, Locust Ridge, Ned Power, Criterion Wind Park, Mountaineer, Roth Rock, and Mt Peak Utility sites.

Data preprocessing is now completed, and we selected Humboldt as the independent variable and Bear Creek as the most important explanatory variable among the nine variables obtained after applying the correlation matrix. Hence, the study was conducted on the non-Bayesian and Bayesian regression models based on the selected data.

B. Non-Bayesian regression models

The present study applies the non-Bayesian model, which is described as a linear regression model. In the model, we use the explanatory variables (Bear Creek, Frey Farm, Locust Ridge, Ned Power, Criterion Wind Park, Mountaineer, Roth Rock, and Mt Peak Utility) to predict the outcome of the response variable (Humboldt). The linear regression model is specified as

$$H_t = a + \beta_1 BC_t + \beta_2 FF_t + \beta_3 LR_t + \beta_4 NP_t + \beta_5 CWP_t + \beta_6 M_t + \beta_7 RR_t + \beta_8 MPU_t + \varepsilon_t, \quad (7)$$

where at the time t, H_t is Humboldt, BC_t is Bear Creek, FF_t is Frey Farm, LR_t is Locust Ridge, NP_t is Ned Power, CWP_t is Criterion, M_t is Mountaineer t, RR_t is Roth Rock, and MPU_t is Mt Peak Utility. The ε_t is the error term, a is the intercept, and β_1, β_2, β_3, β_4, β_5, β_6, β_7, and β_8 are the respective coefficients of the predictive variables.

To determine the fitted model, we applied the variance inflation factor (VIF) on the different linear regression models to measure the amount of multicollinearity. Mathematically, the VIF is the ratio of the overall model variance to the variance of a model that includes only that single independent variable. The ratio is calculated for each independent variable. The high VIF indicates that the associated independent variable is highly correlated with the other variables in the model.

After the imposition of the VIF to the different models, we finally came up with a fitted model that only includes Bear Creek and Mt Peak Utility as the selected explanatory variables, and then we applied the Bayesian approach.

C. Bayesian regression models

This research also applies the Bayesian multiple linear regression. The model assumes that a specific observation has a mean u_t for the ith response variable H_t specified as[18]

$$H_i|u_i, \sigma \sim ind\, Normal(u_i, \sigma), \quad i = 1, \ldots, n, \quad (8)$$

where $i = 1, \ldots, n$, where $n = 8760$ is the number of wind speed observations. This equation shows that each response of the dependent variable independently (ind) follows the normal density function. The standard deviation σ is also distributed among all responses.

The Bayesian multiple regression models is expressed as

$$\mu_i = a + \beta_1 BC_i + \beta_8 MPU_i. \quad (9)$$

The slope parameters β can be interpreted as the change in the expected response u_i when one predictor increases by one unit if the other predictors stay constant.

To implement the Bayesian model, we assign a prior distribution that can have an impact on the posterior distribution. If we assume or believe that the coefficients a and the β_s are independent of σ, then the joint prior density for the coefficients, including the parameter σ, will be written as[18]

$$\pi(a, \beta_1, \beta_8, \sigma) = \pi(a, \beta_1, \beta_8)\pi(\sigma). \quad (10)$$

It is essential to use normal priors by assuming that

$$a \sim Normal(\mu_0, s_0), \quad \beta_1 \sim Normal(\mu_1, s_1), \quad \beta_8 \sim Normal(\mu_8, s_8), \quad (11)$$

where s_j is the standard deviation in the normal prior and shows how we can believe in a prior of β_j. For the prior on sampling the standard deviation σ, we assume Eq. (7), and σ represents the variability of Humboldt wind speed data about the regression line. A Gamma prior for the precision parameter with small values of the shape and rate parameters are written as[18]

$$\phi = \frac{1}{\sigma^2} \sim Gamma(1, 1). \quad (12)$$

In our sampling model, $H_1, \ldots, H_n$ are assumed to be independent with Eq. (7). Suppose that we have a sampling model in which $Y, \ldots, Y_n$ are independent with $Y_i \sim Normal(u_i, \sigma)$ *with* $u_i = \beta_o + \beta_i x_i$ with β_o, β_i, σ representing the parameters. If we assume that $(x_1, y_1), \ldots, (x_n, y_n)$ are observed, the likelihood in the cases will be the joint density of the observations described as a function of $(\beta_o, \beta_1, \sigma)$. The likelihood function is specified as[20]

$$L(\beta_o, \beta_i, \phi) = \prod_{i=1}^{n} \left[\frac{\sqrt{\phi}}{\sqrt{2\pi}} \exp\left\{ -\frac{\phi}{2}(y_i - \beta_o - \beta_i x_i)^2 \right\} \right] \propto \phi^{\frac{n}{2}} \exp\left\{ -\frac{\phi}{2} \sum_{i=1}^{n} (y_i - \beta_o - \beta_i x_i)^2 \right\} \quad (13)$$

The posterior is determined by the multiplication of the prior and the likelihood. The posterior density expression is written as[20]

TABLE II. Multiple regression results. Bold values mean that variables are significant or impacted the dependent variable since their P-Value is less than 0.05.

	Humboldt			Humboldt			Humboldt		
Predictors	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	0.05	−0.04–0.15	0.284	0.44	0.34–0.54	**<0.001**	0.67	0.56–0.78	**<0.001**
Ned Power	0.04	0.02–0.06	**0.001**						
Criterion WP	0.07	0.05–0.09	**<0.001**						
Roth Rock	0.03	0.01–0.06	**0.004**						
Mountaineer	0.01	−0.00–0.02	0.051						
Frey Farm	0.20	0.18–0.21	**<0.001**	0.29	0.27–0.30	**<0.001**			
Locust Ridge	0.08	0.06–0.09	**<0.001**	0.06	0.04–0.07	**<0.001**			
Bear Creek	0.59	0.58–0.61	**<0.001**	0.64	0.62–0.65	**<0.001**	0.87	0.86–0.87	**<0.001**
Mt Peak Utility	0.04	0.03–0.04	**<0.001**	0.05	0.04–0.06	**<0.001**	0.05	0.04–0.06	**<0.001**
Observations		8760			8760			8760	
R^2/R^2 adjusted		0.838/0.838			0.819/0.819			0.780/0.780	

$$\pi(\beta_0, \beta_i, \phi|y_1, \ldots, y_n) \propto \phi^{\frac{n}{2}} \exp\left\{-\frac{\phi}{2}\sum_{i=1}^{n}(y_i - \beta_o - \beta_i x_i)^2\right\} * \exp\left\{-\frac{1}{2s_0^2}(\beta_o - \mu_0)^2\right\} \times \exp\left\{-\frac{1}{2s_i^2}(\beta_i - \mu_i)^2\right\} * \phi^{a-1}\exp(-b\phi). \quad (14)$$

TABLE III. Variance inflation factor.

Dependent variable: Humboldt			
Independent Variables	VIF 1	VIF 2	VIF 3
Ned Power	9.24		
Criterion WP	6.70		
Roth Rock	8.85		
Mountaineer	2.79		
Frey Farm	2.69	2.32	
Locust Ridge	2.66	2.59	
Bear Creek	2.99	2.88	1.00
Mt Peak Utility	1.04	1.03	1.00

To obtain valid inferences from the posterior, convergence of the Monte Carlo Markov Chain (MCMC) should be assessed. MCMC methods are used to approximate the posterior distribution of a parameter of interest by random sampling in a probabilistic space.[20]

D. Kolmogorov–Smirnov goodness of fit test

In this study, the Kolmogorov–Smirnov (K–S) test was applied for hypothesis testing. It is based on the empirical distribution function (ECDF). Given N-ordered data points $Y_1, Y_2, \ldots, Y_N$, the ECDF is defined as[21]

$$E_N = n(i)/N, \quad (15)$$

where $n(i)$ is the number of points less than Y_i and the Y_i are ordered from smallest to largest value.

This step-function increases by $1/N$ at the value of each ordered data point. The K–S test quantifies a distance between the cumulative distribution function of the given reference distribution and the empirical distribution of the given two samples. The K–S Test is only applied to continuous distribution. The test is defined by the following hypothesis:

H_0: The data follow a specified distribution
H_a: The data do not follow the specified distribution

The K-S Test statistics is defined as[21]

TABLE IV. Multiple linear regression results from a different sample size. Bold values mean that variables are significant or impacted the independent variables. When P-value is less than 0.05, we reject the null Hypothesis.

	Humboldt			Humboldt			Humboldt		
Predictors	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	0.67	0.56–0.78	**<0.001**	−0.03	−0.39–0.33	0.851	−0.12	−1.30–1.07	0.843
Bear Creek	0.87	0.86–0.87	**<0.001**	0.95	0.92–0.98	**<0.001**	0.95	0.83–1.06	**<0.001**
Mt Peak Utility	0.05	0.04–0.06	**<0.001**	0.05	0.02–0.09	**0.002**	0.08	−0.07–0.22	0.297
Observations		8760			365			53	
R^2/R^2 adjusted		0.780/0.780			0.912/0.911			0.869/0.863	

$$D_n = max_{1\le i\le N}\left(F\left(Y_i - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)\right)\right), \quad (16)$$

where F is the continuous theoretical cumulative distribution being tested.

Another measurement of the test is the *P*-value. The hypothesis of the distributional form is rejected if the D value is greater than the *P*-value. Another explanation of the K–S test is its *P. value.* If its *P*-value is less than the significant level (0.05), we reject the null hypothesis that the two samples were drawn from the same distribution.

V. RESULTS

A. Multiple linear regression model results

In this study, the first inferential analysis implemented using the wind speed data is the multiple linear regression (non-Bayesian). Equation (7) is used to evaluate the impact of the predictors on the dependent variable. All eight predictors are first incorporated into the model. After getting the first results from the model (Table II), VIF was applied for the multi-collinearity effect among the predictors.

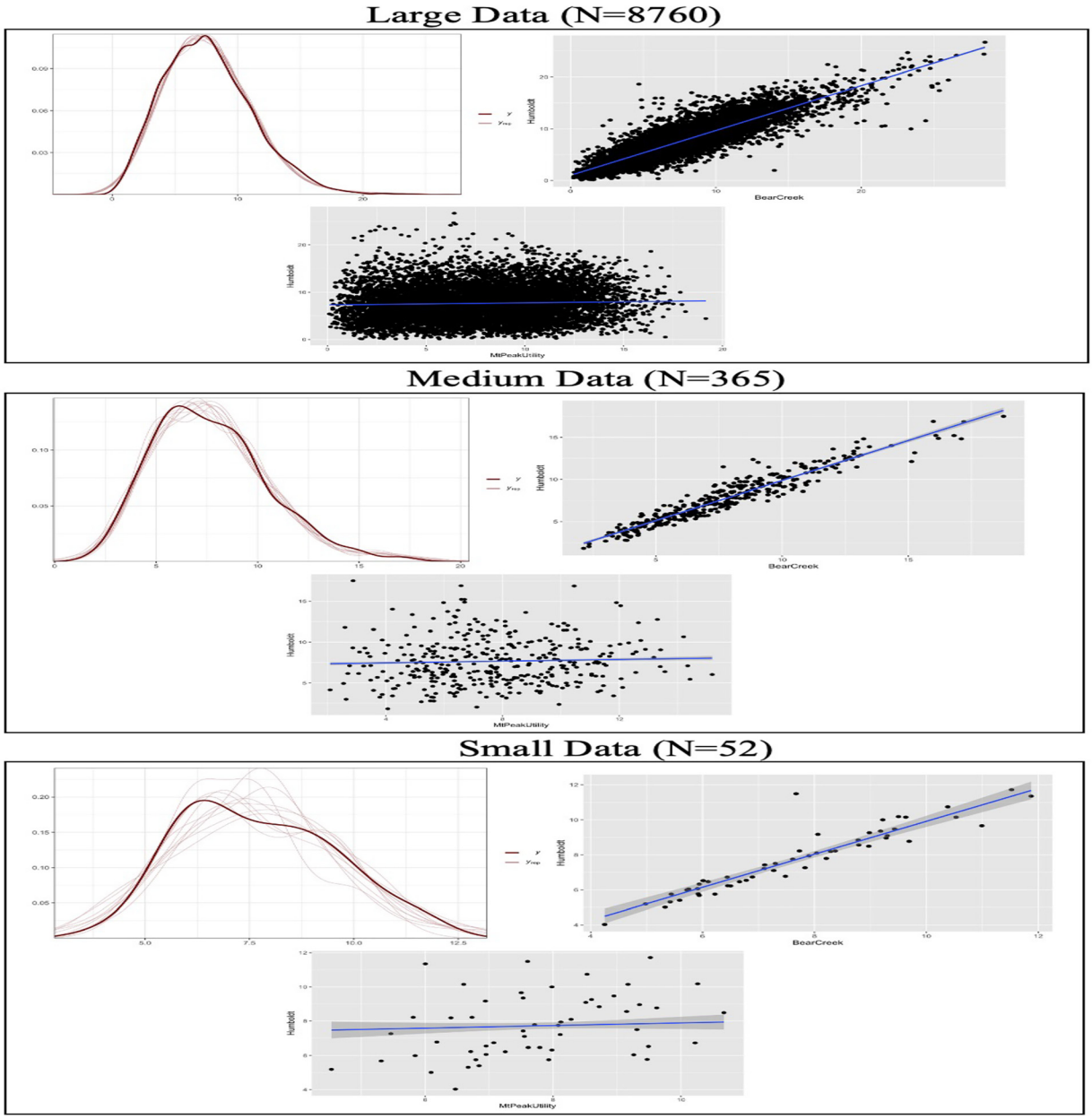

FIG. 3. Posterior prediction results.

Table III depicts that three of the variables (Ned Power, Roth Rock, and Criterion Wind Park) have their VIFs respective values of 9.24, 8.85, and 6.70, which are large. For example, the VIF of Ned Power is very high and is highly correlated with at least one of the other predictors. The same interpretation is also applied to Roth Rock and Criterion Wind Park. Mountaineer is not significant to the model after running Eq. (7). After removing the variables with a high variance inflator factor, we run a second model that only includes Frey Farm, Locust Ridge, Beak Creek, and Mt Peak Utility. As a result, all predictors have shown a significant impact on Humboldt, but the problem of multicollinearity among the predictors was still present. VIF was once again applied to the model and we found that Frey Farm, Locust Ridge, and Bear Creek have a respective value of 2.32, 2.66, and 2.99, but as stated earlier, Bear Creek was kept in the model because it is our variable of interest based on its important impact on the dependent variable.

Finally, the obtained fitted model that is used for the rest of the study is expressed as

$$\begin{aligned} H_t &= a + \beta_1 BC_t + \beta_2 MPU_t + \varepsilon_t \\ \widehat{H_t} &= 0.67 + 0.87 BC_t + 0.05 WPU_t. \end{aligned} \tag{17}$$

$\widehat{H_t}$ is defined as the predicted value of H_t (Humboldt) in regression Eq. (17).

In this model, the remaining VIF is very satisfactory. Both Bear Creek and Mt Peak Utility, respectively, have 1.00 and 1.00, which means there is no multi-correlation. From the result, both predictors are statistically significant because of their respective *P*-values which are less than 0.05.

From the fitted model, we see that a one-unit shift in Bear Creek increases Humboldt by 0.87 on average while holding Mt Peak Utility as a constant. Same for Mt Peak Utility which increases Humboldt on average by 0.05 while Bear Creek remains constant.

This study also considered the sensitivity analysis to measure the uncertainties associated with the wind speed data sample sizes (hourly with 8760 observations, daily with 365 observations, and weekly data with 53 observations). We then run the multiple linear regression using Eq. (17). As shown in Table IV, Bear Creek's coefficient is positive and statistically significant in all regression results and pretty high in the daily and weekly data sample sizes with a value of 0.95 in both. Mt Peak Utility, instead, is positive and statistically significant in only two regression results, respectively, for the hourly and daily datasets. Additionally, the R^2 value (0.91) from the daily data is pretty high compared to that of hourly data (0.78) and weekly data (0.86).

This value of %91 indicates that the model explains all the variability of the response data around its mean when using the daily data. Figure 3 shows less uncertainty on the daily data.

1. Bayesian and MCMC regression results

The second inferential model that we have explored in our study is the Bayesian multiple regression model. As a part of the Bayesian inference process, priors must be defined for the regression coefficients.

From Table V, the regression parameters are alike to the ones we obtained from our non-Bayesian model where priors were not constructed. The marginal posterior probability for Bear Creek and Mt Peak is around 0.87 and 0.05, respectively.

Three priors were used for all three datasets: an uninformative gamma prior, a uniform prior, and an informative normal prior. This was done to examine the effect of priors and observation sizes on the estimation of regression coefficients. It is expected that the estimated regression coefficients would deviate from the non-Bayesian estimates using the gamma prior. Moreover, the weekly dataset with the least number of observations is expected to show more uncertainty in estimation.

As shown in Fig. 4, the choice of prior has more effect on the smaller dataset than the larger ones in coefficient estimation. When there are many observations, the priors do not have a significant effect on the estimation of regression coefficients. Regardless of the prior, the estimated regression parameters stay within a certain range of values close to the non-Bayesian coefficients.

TABLE V. Bayesian regression model.

Bayesian regression model				
Deviation Residuals:				
Min	1Q	Median	3Q	Max
−11.149	−1.0214	−0.1376	0.9528	13.1351
Coefficients:				
	Estimate	Std. error	t value	Pr(>|t|)
(Intercept)	0.671 201	0.056 393	11.902	$<2 \times 10^{-16}$***
Bear Creek	0.865 158	0.004 938	175.194	$<2 \times 10^{-16}$***
Mt Peak Utility	0.045 658	0.005 166	8.838	$<2 \times 10^{-16}$***
Signif. codes	0 "***" 0.001 "**"	0.01"*"	0.05 "." 0.1 " "	1
(Dispersion parameter for Gaussian family taken to be 2.826 083)				
Null deviance: 112 482 on 8759 degrees of freedom				
Residual deviance: 24 748 on 8757 degrees of freedom				
AIC: 33 965				
Number of Fisher scoring iterations: 4				

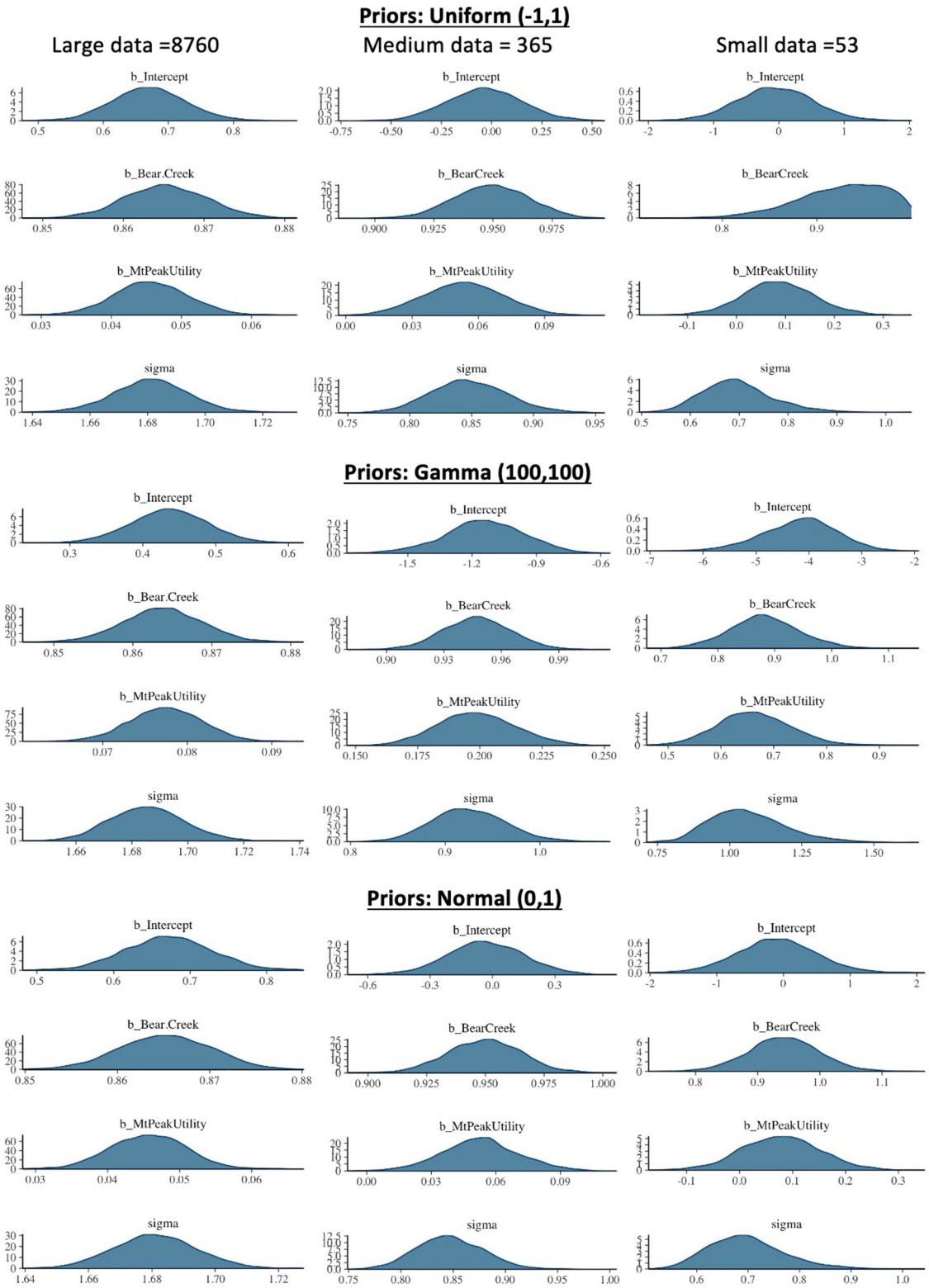

FIG. 4. Priors' effects on different wind speed data size.

TABLE VI. MCMC regression model results.

MCMC regression model

Iterations = 1001:2000
Thinking interval = 1
Number of chains = 1
Sample size per chain = 1000
1. Empirical mean and standard deviation for each variable plus standard error of the mean:

	Mean	SD	Naive SE	Time - series SE
(Intercept)	0.6754	0.055 516	0.00 017 556	0.0 015 056
Bear Creek	0.8649	0.004 737	0.0001498	0.0 001 498
Mt Peak Utility	0.0455	0.005 229	0.0 001 653	0.0 001 511
Sigma2	2.8262	0.043 548	0.0 013 771	0.0 013 771

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
(Intercept)	0.56 712	0.63 751	0.67 503	0.7134	0.78 038
Bear Creek	0.85 498	0.86 171	0.86 494	0.8681	0.87 383
Mt Peak Utility	0.03 538	0.04 199	0.04 562	0.049	0.05 596
Sigma2	2.73 911	2.79 707	2.82 646	2.8573	2.91 005

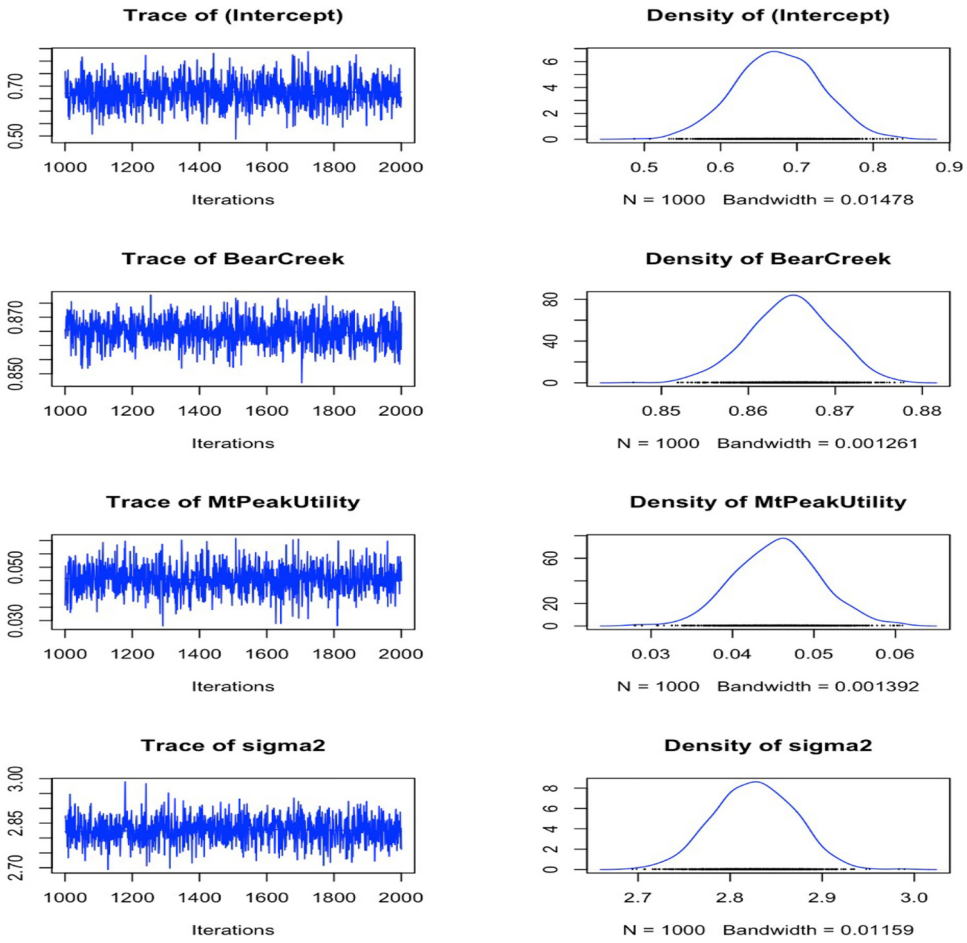

FIG. 5. Trace plot of the posterior distribution.

TABLE VII. K–S Test results.

K–S test results		
Wind speed data	D	P-Value
Humboldt/Bear Creek	0.0141	0.344
Humboldt/Mt Peak Utility	0.0699	$< 2.2 \times 10^{-16}$

Our Bayesian model is followed by a diagnostic Monte Carlo Markov Chain (MCMC) simulation. From the MCMC, we obtained valid inferences from the posterior. We used the default interactions of 2000. When expecting the location of the 90% probability interval, the interval for the estimated parameter corresponding to Bear Creek is (0.8549, 0.87383) and the corresponding estimate for Mt Peak Utility is (0.035, 0.055). Table VI indicates that both Bear Creek and Mt Utility Peak help to predict Humboldt. The diagnostic results obtained from the density plot providing the view of the posterior distribution of the parameters are well seen in Fig. 5. For example, the mean of the posterior distribution of Bear Creek lies between 0.86 and 0.87.

2. Kolmogorov–Smirnov results

We used the Kolmogorov–Smirnov test in this research to see if the two groups of selected wind speed data variables, such as Humboldt vs Bear Creek, and Mt peak Utility vs Bear Creek, were sampled from populations with different distributions. This test is designed to evaluate if there is a difference in median, variability, or the shape of the distribution.

From Table VII, the *P*-values resulted from the respective Humboldt and Bear Creek distribution are greater than the significant value (0.05), which means that the population may not differ in the median, variability, or the shape of their distribution. The case is not seen with Humboldt and Mt Peak Utility where the *P*-values are lower than the significant *P*-value (0.05).

We graphically used the empirical cumulative distribution functions (ECDF) (see Fig. 6) to analyze the distribution of the data. This further validates our interpretation of the distribution of Humboldt and Bear Creek. The distributions are the same and the distance from the two distributions is not visible between Humboldt and Bear Creek.

VI. CONCLUSION

Statistical methods are widely used to analyze the relationship between several independent variables (predictors) and a dependent variable. As wind energy rapidly becomes an important source of renewable energy, it is very important to deeply evaluate any potential existing relationships among the data. This paper aimed to apply the non-Bayesian and the Bayesian approaches to multiple linear regressions for wind speed data and investigated the differences between the two statistical methods. The NREL wind speed data were used for fifteen different wind farms. In the proposed study, a correlation matrix was implemented to select the most correlated variables and determine the highest correlated variable among them, and used it as the

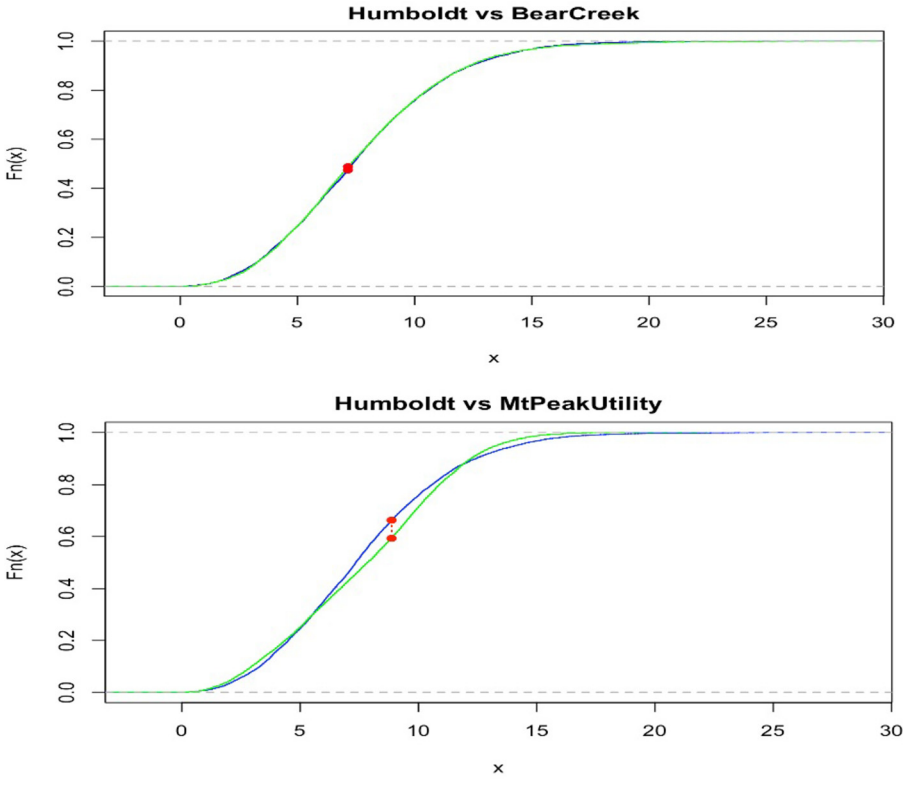

FIG. 6. Empirical cumulative distribution function graph.

dependent variable. This method was followed by a RF machine learning technique for the feature selection and considering the most important features that would be used for the non-Bayesian and Bayesian regression models. We first ran a multiple linear regression on the non-Bayesian regression model in which we applied the VIF to detect the multicollinearity problem to get the fitted model. We then applied the Bayesian approach to the fitted model to analyze the relationship between the dependent and independent variables. Even though several studies based on the Bayesian development models have shown better results than the non-Bayesian methods, our study showed both the non-Bayesian and the Bayesian approaches are very much alike in the coefficients/parameters estimations. Moreover, we analyzed data with different sample sizes (hourly, daily, and weekly), and the daily data provided a strong coefficient estimator and highest R-squared compared to the hourly and weekly datasets.

Regardless of the prior, the estimated regression parameters stayed within a certain range close to the non-Bayesian coefficients. In the future, we will implement vine copula approaches, such as R, C, and D-vines, to analyze the high-dimensional dependency modeling in wind speed data. We will also investigate the vine copula model uncertainties using the Bayesian vine copula approaches. Eventually, the proposed non-Bayesian and Bayesian models will be examined on the famous power systems' problems, such as DC and AC optimal power flow, stochastic unit commitment, stochastic economic dispatch, and power systems' resiliency.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant No. 1900462.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]B. D. Cross, K. E. Kohfeld, J. Bailey, and A. B. Cooper, "The impact of wind speed trends and 30-year variability in relation to hydroelectric reservoir inflows on wind power in the Pacific Northwest," PloS One **10**(8), e0135730 (2015).

[2]Z. Wang, W. Wang, C. Liu, Z. Wang, and Y. Hou, "Probabilistic forecast for multiple wind farms based on regular vine copulas," IEEE Trans. Power Syst. **33**(1), 578–589 (2017).

[3]S. Gill, B. Stephen, and S. Galloway, "Wind turbine condition assessment through power curve copula modeling," IEEE Trans. Sustainable Energy **3**(1), 94–101 (2011).

[4]Q. Xu, Z. Fan, W. Jia, and C. Jiang, "Fault detection of wind turbines via multivariate process monitoring based on vine copulas," Renewable Energy **161**, 939–955 (2020).

[5]W. Zhaol, Y. Fu, Z. Zheng, B. Chen, Q. Liao, W. Xie, and B. Yang, "Correlation analysis of wind power based on mixed copula model and its application into stochastic dispatch," in *International Conference on Power System Technology (POWERCON)* (IEEE, 2018), pp. 1062–1069.

[6]M. S. Li, Z. J. Lin, T. Y. Ji, and Q. H. Wu, "Dispatch considering dependence of multiple wind farms using paircopula," Appl. Energy **226**, 967–978 (2018).

[7]M. A. Mohamed, A. Hajjiah, K. A. Alnowibet, A. F. Alrasheedi, E. M. Awwad, and S. M. Muyeen, "A secured advanced management architecture in peer-to-peer energy trading for multi-microgrid in the stochastic environment," IEEE Access **9**, 92083–92100 (2021).

[8]H. Zou, J. Tao, S. K. Elsayed, E. E. Elattar, A. Almalaq, and M. A. Mohamed, "Stochastic multi-carrier energy management in the smart islands using reinforcement learning and unscented transform," Int. J. Electr. Power Energy Syst. **130**, 106988 (2021).

[9]L. Min, K. A. Alnowibet, A. F. Alrasheedi, F. Moazzen, E. M. Awwad, and M. A. Mohamed, "A stochastic machine learning based approach for observability enhancement of automated smart grids," Sustainable Cities Soc. **72**, 103071 (2021).

[10]M. A. Mohamed, T. Jin, and W. Su, "An effective stochastic framework for smart coordinated operation of wind park and energy storage unit," Appl. Energy **272**, 115228 (2020).

[11]O. Seidou, J. J. Asselin, and T. B. M. J. Ouarda, "Bayesian multivariate linear regression with application to changepoint models in hydrometeorological variables," Water Resour. Res **43**, W08401 (2007).

[12]See https://data.nrel.gov/submissions for "data used in the paper."

[13]D. Syarifah and T. Heruna, "Linear regression model using Bayesian approach for energy performance of residential building," Elsevier Procedia Comput. Sci. **135**, 671–677 (2018).

[14]D. Birkes and Y. Dogde, *Alternative Methods of Regression* (John Wiley and Sons, Inc, 1993).

[15]X.-W. Ye, Y. Ding, and H.-P. Wan, "Probabilistic forecast of wind speed based on Bayesian emulator using monitoring data," Struct Control Health Monit. **28**, 2650 (2021).

[16]T. K. Ho. "Random decision forests (PDF)," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, Quebec (Tin Kam Ho, 1995) pp. 14–16.

[17]R. Stacey, see https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3 for "The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-Learn and Spark, Toward Data Science" (last accessed May 11, 2018).

[18]J. Albert and J. Hu, *Probability and Bayesian Modeling* (Chapman and Hall/CRC, 2019), Chap. 12–12.2.

[19]I. M. Chakravarti, R. G. Laha, and J. Roy, *Handbook of Methods of Applied Statistics* (John Wiley and Sons, 1967), Vol. 1, pp.392–394.

[20]J. Albert and J. Hu, *Probability and Bayesian Modeling* (Chapman and Hall/CRC, 2019), Chap. 10–10.2.3.

[21]NIST SEMATECH, see http://www.itl.nist.gov/div898/handbook, for "e-Handbook of Statistical Methods."