Performance Analysis of Short and Mid-Term Wind Power Prediction using ARIMA and Hybrid Models.

Ashoke Kumar Biswas,
Sina Ibne Ahmed,
Temitope Bankefa,
Prakash Ranganathan, *Senior Member, IEEE*
Hossein Salehfar, *Senior Member, IEEE*
School of Electrical Engineering and Computer Science
University of North Dakota, Grand Forks, U.S.A.
ashoke.biswas@und.edu, sina.ahmed@und.edu,
temitope.bankefa@und.edu, prakash.ranganathan@und.edu, h.salehfar@und.edu

2021 IEEE Power and Energy Conference at Illinois (PECI) | 978-1-7281-8648-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/PECI51586.2021.9435209

Abstract— **Due to the high market penetration of wind power, efficient prediction methodologies are of paramount importance to promote wind power generation in the electricity market against more secure and dispatchable energy sources. This paper has proposed mix of regression and machine learning methods, such as Auto-Regressive Integrated Moving Average (ARIMA), Random Forest (RF), Bagging Classification and Regression Trees (BCART), and two hybrid models of ARIMA-RF and ARIMA-BCART to forecast one, two, and seven days of wind power generation. The prediction relies on weather data such as wind speed, wind direction, air temperature, air pressure, and density at hub height. The preliminary results indicate that ARIMA-RF and ARIMA-BCART aids in improving forecasting accuracy (i.e., NMAE 18%-26%) over standalone forecast mode of ARIMA.**

Key words: Wind Power Forecasting, ARIMA, Random Forest (RF), BCART, ARIMA-RF. ARIMA-BCART.

I. Introduction

Many countries have initiated steps to reduce carbon emissions by deploying more renewable energy sources, especially wind energy. The U.S. Federal Government has granted new energy and environmental policies and objectives to expand renewable energy use significantly [1]. Global energy demand will rise 40% by 2040, and fossil fuels will offset two-thirds of the total demand [2]. The need for a robust interface of renewable energy into the power grid has been highlighted by energy and environmental concerns. A model has been developed by the U.S. Department of Energy (DOE), where 20% of U.S. electricity demand will be provided by wind energy by 2030 [3].

Large-scale wind power integration will introduce scheduling and dispatch problems for the energy sector and power system operators, which require mitigation. Wind Power Forecasting (WPF), one of the least expensive and most straightforward approaches, can be used to reduce the negative impacts of intermittent wind generation [4]. Wind power generation is gaining critical importance in the market environment, since forecasting will impact electricity prices [1]. Therefore, a reliable prediction model is critical for the smooth operation of wind power.

ARIMA is a popular method for time series forecasting that has temporal trend and seasonality data for input features [5]. It uses historical data to capture linear relationships among the time periods. The ARIMA model alone is insufficient for dealing with non-linear data, whereas machine learning models are not equally capable of managing both linear and non-linear data [6]. The machine learning methods of Random Forest (RF) and Bagging Classification and Regression Trees (BCART) improve the prediction accuracy by capturing the non-linear features of the data. In this paper, the authors propose hybrid model to improve the prediction performance of the ARIMA model for predicting wind power generation prediction of US East and West coast. The geographical location of US East and West coast are shown in TABLE I.

TABLE I. GEOGRAPHICAL LOCATION OF US EAST AND WEST COAST

Site	Location	
	Longitude	Latitude
East Coast-1	-69.4597	40.86165
East Coast-2	-69.483	40.86165
West Coast-1	-124.587	42.80292
West Coast-2	-124.596	42.82035

This paper has investigated three single-stage models of ARIMA, RF, BCART, and two hybrid models of ARIMA-RF and ARIMA-BCART to compare prediction accuracy among individual models, and to observe prediction performance improvement of ARIMA by using hybrid model. This paper has also analyzed the performance of the prediction models using the three significant weather variables. Two hybrid models of ARIMA-RF and ARIMA-BCART have been suggested for precise prediction. The performance of the forecasting methods for short and medium-term wind power generation of US East and West coast has also been discussed in this paper.

II. Single –Stage Models

A. Non-Seasonal ARIMA Model

ARIMA is one of the most sought-after stochastic models for analyzing time-series data and predicting wind power generation. ARIMA comprises different types of time series, such as pure autoregressive (AR), integrated (I), pure moving average (MA), and combined AR and MA (ARMA) models [5], [6]. The different components of these time series are applied for filtering the respective residuals denoted by p, d, and q. The trend information must first be extracted, and the differencing operation is used to create the residual stationary or zero mean series. Once the time-series data is stationary, the autoregressive steps activates and determines the forecast value compared to the present value [5]. An autoregressive model of order p, abbreviated AR (p), is illustrated in eq. (1) [5], [7].

$$y_t = \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots. + \emptyset_p y_{t-p} + \epsilon_t \quad (1)$$

Where y_t is stationary and represents power generation at time t, and $\emptyset_1, \emptyset_2, \ldots, \emptyset_p$ are regression coefficients ($\emptyset_p \neq 0$). ϵ_t is the error term, which is Gaussian white noise with mean zero. The moving average model of order q or the MA (q) model is defined as eq. (2) [7].

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots. + \theta_q \epsilon_{t-q} \quad (2)$$

The p and q variables are known as AR and MA orders, respectively. The generalization of the ARMA model is represented by eq. (3) [7].

$$y_t = \alpha + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots. + \emptyset_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots. + \theta_q \epsilon_{t-q} \quad (3)$$

The ARIMA (p, d, q) is illustrated by eq. (4) [7]

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t \quad (4)$$

Where, $\nabla^d y_t = (1-B)^d y_t$ is ARMA (p, q).

B. Seasonal ARIMA Model

The regular periodic patterns in the time series are known as seasonality, denoted by parameter S. ARIMA has been used extensively to predict seasonal time series. The multiplicative seasonal ARIMA model is illustrated as [2]:

$$ARIMA\ (p,d,q) * (P,D,Q)s \quad (5)$$

Where
p = order of non-seasonal AR terms,
P = order of seasonal AR terms,
q = order of non-seasonal MA terms,
Q = order of seasonal MA terms,
d = order of non-seasonal differencing,
D = order of seasonal differencing,
And s = span of seasonality pattern.

The seasonal ARIMA or SARIMA model is expressed as the following eq. (6) [7].

$$\Phi_P(B^s)\phi(B)\nabla_s^D \nabla^d y_t = \delta + \Theta_Q(B^s)\theta(B)\epsilon_t \quad (6)$$

Where, ϵ_t is the error term.
Here,
$\Phi_P(B^s) = (1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps})$,
$\Theta_Q(B^s) = (1 + \Theta_1 B + \Theta_2 B^{2s} \ldots + \Theta_Q B^{Qs})$,
$\phi(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 \ldots - \emptyset_P B^p$,
$\theta(B) = 1 + \theta_1 B + \theta_1 B^2 + \cdots + \theta_Q B^q)$,
$\nabla^d = (1-B)^d$, $\nabla_s^D = (1-B^s)^D$.

The final equation is denoted with eq. (7).

$$(1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots. - \emptyset_P B^p)(1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps})\,(1-B^s)^D (1-B)^d y_t = \big(1 + \theta_1 B + \theta_1 B^2 + \cdots + \theta_Q B^q)\big)(1 + \Theta_1 B + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs})\epsilon_t \quad (7)$$

The non-seasonal AR and MA parts are denoted by $\phi(B)$ and $\theta(B)$ with orders p and q, respectively, and the seasonal AR and MA parts are denoted by $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ of orders P and Q.

C. Random Forest Model:

A decision tree was presented by Breiman in 1984 and Random Forest, the generalization of decision trees, was presented by Breiman in 2001[8], [9], [10]. This method aggregates tress and is used for the classification or regression to avoid overfitting. Among the tree predictors, the most important tree is voted for in the forest and among the large number of trees [11]. The error for forests becomes low if the number of trees is large. This method can handle a vast number of features and assist in choosing features based on significance. This model is user friendly and only uses two free parameters of bootstrapping ensembles n_{trees} (default 500), where n is the number of trees, and the randomized input predictors sample is m_{try} (default 2) [9].

Random Forest has an ensemble of *B* trees $\{T_1(X), \ldots, T_B(X)\}$, where $X = \{x_1, \ldots, x_p\}$ is a *p*-dimensional vector of explanatory variables. The ensemble generates B outputs $\{\hat{Y}_1 = T_1(X), \ldots, \hat{Y}_B = T_B(X)\}$,where $\hat{Y}_b = 1, \ldots, B$, is the prediction of the *b*th tree [12]. The final prediction is the average of the individual tree prediction, $\hat{Y}$ [10], [11], [12]. The training method involves constructing a predictor, h(X), where the characteristics are recursively divided into nodes with distinct levels, Y. The RF, ensemble-based approach only emphasizes the ensemble of decision trees and gives this machine learning approach flexibility and computing power [5]. This requirement is not possible when children nodes with different labels exist. The terminal nodes are referred to as tree leaves and display the various possible labels of Y [10]. If the tree predictor of a random forest is *h(X)* and the distribution of vector Y, X, the mean-squared error for any numerical predictor is [11]

$$E_{X,Y}\big(Y - h(X)\big)^2$$

D. Bagging Classification and Regression Trees (BCART) Model:

Bootstrap Aggregating (bagging) is a widely used technique to combine many predictors to create a precise technique, introduced in 1994 by Leo Breiman [13]. This technique uses the bootstrap replication method to train the original data set, and a predictor is produced for each replicate sample. The predictors are combined using the average function for regression, and the majority vote for classification [13].

Given a learning set $\mathcal{L} = \{(y_n, x_n), n = 1, \ldots., N\}$, where, y is the class level, or a result. If x is an input, the output predictor, y, is denoted $\varphi(x, \mathcal{L})$. If the output y is the numerical response, the average function of the predictor is proposed by Breiman is [13], [14]:

$$\varphi_B(x) = av_B \varphi\big(a, \mathcal{L}^{(k)}\big) \quad (8)$$

1. Bagging Classification Trees

The data set is organized indiscriminately into a test set T and learning set $\mathcal{L}$, and in most cases, $\mathcal{L}$, would be reasonably large [14]. The 10-fold cross-validation builds a classification tree from $\mathcal{L}$. This tree generates the miscalculation rate $e_s(\mathcal{L}, \mathrm{T})$. Using 10-fold cross-validation, a tree is built using a bootstrap sample, $\mathcal{L}_B$.

The tree classifiers will be $\varphi_1(x), \varphi_2(x), \ldots\ldots, \varphi_{50}(x)$, if it repeats 50 times. If $(j_n, x_n)(x) \in \mathrm{T}$, x_n has the plurality of $\varphi_1(x_n), \varphi_2(x_n), \ldots\ldots, \varphi_{50}(x_n)$. If the estimated class differs from the original, the bagging miscalculation rate is $e_B(\mathcal{L}, \mathrm{T})$.

2. Bagging Regression Tress

The data set is processed spontaneously into a test set T and learning set $\mathcal{L}$. Normally, an $\mathcal{L}$ of 200 cases is generated for the learning set, and 1,000 cases are generated for the test set. By 10-fold cross-validation, a regression tree is built from $\mathcal{L}$. The tree creates the mean-squared-error of $e_s(\mathcal{L}, \mathrm{T})$ [14]. A regression tree is built using a bootstrap replicate $\mathcal{L}_B$. The predictor will be $\varphi_1(x), \varphi_2(x), \ldots\ldots, \varphi_{25}(x)$, if it repeats 50 times. If $(y_n, x_n)(x) \in \mathrm{T}$, the predicted $\hat{y}_B$ value will be $av_k\varphi_k(x_n)$. The mean-squared-error is $e_B(\mathcal{L}, \mathrm{T})$ in T. The single tree and bagged error over 100 iterations are $\overline{e}_s$ and $\overline{e}_B$.

III. Hybrid Models

Real-world time series data are rarely purely linear or non-linear, although both linear and non-linear data are typically used. The ARIMA model alone is not sufficient for non-linear data management, while machine learning models are not equally capable of managing both linear and non-linear data [6]; therefore, no single approach is appropriate [6], [15]. The Monte Carlo simulation, or bootstrapping method, has been popularized to forecast non-linear patterns because the distribution of the error mechanism does not require any assumptions [16]. The machine learning RF model uses bootstrap sampling, and the bootstrap replication method is used in the BCART model. Compared to a single process, the hybrid method performs well [17]. We can capture various aspects of the underlying trends of time series results by integrating differences and can capture the underlying patterns of time series data by combining different models. If we assume that a time-series data set consists of a linear autocorrelation structure and a non-linear component, the data should be as follows [6]:

$$y_t = L_t + N_t \tag{9}$$

Where the linear component is denoted by L_t and the non-linear component is denoted by N_t. For the linear component, we use ARIMA and then determine the residuals from the linear component that contains a non-linear relationship. The residual from the linear model at time t is illustrated as follows:

$$r_t = y_t - \widehat{L_t} \tag{10}$$

Here, $\widehat{L_t}$ is the forecast value with time t.

By modeling residuals using the bootstrap sampling method in RF and the bootstrap replication method in BCART, the non-linear forecast value $\widehat{N}_t$ can be found. Then, the combined forecasted value will be as follows:

$$\hat{y}_t = \widehat{L_t} + \widehat{N}_t \tag{11}$$

A. ARIMA-RF

The predictor variables are used to train the ARIMA model in the first step. If the relationship between wind power generation and the atmospheric variables is non-linear, ARIMA will not capture the non-linear component of the data; however, the ARIMA model's residual will contain non-linear information. The residuals from the ARIMA model are used to analyze the non-linear structure of the data in the second step, after which we combine the forecasts to improve the overall performance. If the forecasted value in the first step is $\hat{F}$ and the calculated forecasted value from the second step is $\hat{R}$, the final forecasted value will be as follows:

$$\hat{Y} = \hat{F} + \hat{R} \tag{12}$$

The flowchart of the hybrid two-stage model is displayed in Fig. 1 [5].

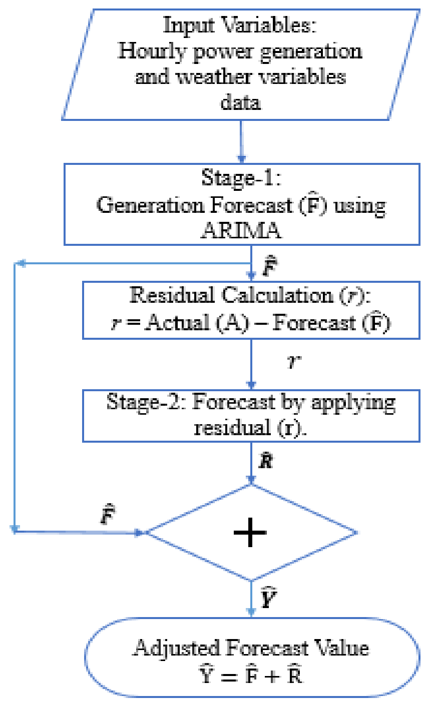

Fig. 1. Two-stage hybrid forecast model.

B. ARIMA-BCART

The initial stage is similar to the previous model. The residuals from the ARIMA model are fed to the BCART model to predict the forecasted value from the residuals. If the forecasted value in the first step is $\hat{F}$ and the calculated forecasted value from the second step is $\hat{R}$, the final forecasted value will be as follows:

$$\hat{Y} = \hat{F} + \hat{R} \tag{13}$$

IV. METHODOLOGY

This paper has used the National Renewable Energy Laboratory (NREL) wind toolkit's last updated data for 2009-2012, which contains estimated wind power generation data from four different sites along the US East and West coasts. The training datasets contain wind turbine power data and weather data such as wind speed, wind direction, air temperature, surface air pressure, and air density at hub height [18]. The models are used to predict wind power for three different durations: 24 hours, 48 hours, and 7 days. The performance of these models has been analyzed by using two commonly used statistical indices, Normalized Mean Absolute Error (NMAE) and Normalized Root Mean Square Error (NRMSE):

$$MAE = \frac{\sum_{i=1}^{N}|P_{ai}-P_{fi}|}{N} \qquad (14)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_{ai}-P_{fi})^2} \qquad (15)$$

where P_{ai} and P_{fi} respectively signify the actual and forecasting value of wind power output at time i and N is the number of forecast samples involved.

$$NMAE = \frac{MAE}{Mean\ of\ the\ Generation\ Data} \times 100\% \qquad (16)$$

$$NRMSE = \frac{RMSE}{Mean\ of\ the\ Generation\ Data} \times 100\% \qquad (17)$$

Root Mean Square Error (RMSE) is the square root of MSE. It is utilized more generally than MSE because MSE value sometimes can be too big to compare easily, and square root brings it back the same level of forecast error. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable. The NMAE and NRMSE has interpreted the error as a fraction of the overall range.

Seasonality has been found among for the datasets of the three sites. Independent variables do not equally contribute to the construction of the prediction model. This paper has used the Spearman correlation method to determine the significant features for each site. Input variables are ranked based on their correlation co-efficient, and shown in TABLE II.

TABLE II. SPEARMAN CORRELATION COEFFICIENT OF INPUT VARIABLES FOR WIND POWER GENERATION.

Explanatory variables	East Coast-1	East Coast-2	West Coast-1	West Coast-2
Wind speed at 100m height	0.922	0.920	0.85	0.844
Wind direction at 100m height	0.102	0.102	-0.30	-0.288
Air temperature at 2m	-0.174	-0.175	0.16	0.147
Surface air pressure	-0.304	-0.302	-0.10	-0.099
Air density	0.045	0.045	-0.11	-0.099

The data in TABLE II illustrates that the three significant weather variables from the East coast data sets are different from the West coast data sets. Wind speed is the most important feature, while wind direction and air temperature are the next two important variables for the west coast sites, and surface air pressure and air temperature are the most important variables for the east coast sites. Insignificant variables can be removed from the training datasets, as their contribution will be minimal for wind power generation prediction. This paper has also compared the performance of the algorithms based on these three most important variables.

V. RESULTS AND DISCUSSIONS

The single-stage method of ARIMA, RF and BCART, and the two-stage hybrid models of ARIMA-RF and RIMA-BCART have been used to predict wind power generation (MW) forecasting from the US East and West coast windfarms. The models are analyzed using the datasets with durations of 24 hours, 48 hours, and 7 days. We have evaluated the accuracy of the forecasted models with the statistical indices of NMAE and NRMSE.

The comparison of the performance in terms of NMAE with the different datasets from the US East and West coasts with five weather variables are shown in TABLE III and Fig. 2; the three significant weather variables are shown in TABLE IV and Fig. 3.

TABLE III. COMPARISON OF NMAE FOR FIVE WEATHER VARIABLES DATA

Time	Location	ARIMA	RF	BCART	ARIMA-RF	ARIMA-BCART
24H	Eastcoast-1	33.63%	0.52%	6.60%	26.52%	26.75%
	Eastcoast-2	33.46%	0.53%	6.56%	30.05%	29.96%
	Westcoast-1	24.29%	0.27%	5.50%	20.70%	21.76%
	Westcoast-2	18.32%	0.19%	4.89%	16.60%	16.47%
48H	Eastcoast-1	33.42%	0.98%	8.20%	27.12%	27.25%
	Eastcoast-2	33.10%	1.29%	9.23%	27.19%	28.34%
	Westcoast-1	26.11%	0.32%	6.05%	22.05%	22.77%
	Westcoast-2	20.33%	0.23%	6.02%	17.47%	17.23%
7 DAYS	Eastcoast-1	29.82%	1.79%	9.49%	24.02%	24.27%
	Eastcoast-2	30.23%	2.18%	9.54%	24.90%	26.81%
	Westcoast-1	38.23%	4.03%	12.19%	31.80%	34.26%
	Westcoast-2	40.46%	3.40%	13.71%	33.59%	35.35%

TABLE IV. COMPARISON OF NMAE FOR THREE SIGNIFICANT WEATHER VARIABLES.

Time	Location	ARIMA	RF	BC	ARIMA-RF	ARIMA-BCART
24H	Eastcoast-1	33.40%	0.50%	7.56%	29.71%	28.71%
	Eastcoast-2	31.40%	0.49%	7.35%	27.26%	27.15%
	Westcoast-1	18.10%	0.37%	5.98%	15.54%	17.23%
	Westcoast-2	18.32%	0.37%	6.12%	16.60%	16.85%
48H	Eastcoast-1	33.36%	0.96%	9.41%	28.13%	26.92%
	Eastcoast-2	32.38%	0.85%	9.31%	27.56%	28.37%
	Westcoast-1	18.50%	0.37%	6.47%	15.82%	17.81%

	Westcoast-2	19.79%	0.37%	6.68%	16.26%	16.97%
7 DAYS	Eastcoast-1	29.97%	1.24%	10.87%	24.03%	26.11%
	Eastcoast-2	29.78%	0.93%	10.66%	24.18%	25.31%
	Westcoast-1	35.24%	5.86%	14.61%	29.20%	30.47%
	Westcoast-2	34.08%	5.51%	14.91%	15.87%	16.47%

The TABLE III and Fig. 2 show, RF model has better prediction rates for all time durations. Since, the relationship between the output and predictor variables are non-linear, and ARIMA is not suited for modeling non-linear data, the performance of ARIMA is the least among all methods; however, RF and BCART have greater prediction accuracy in terms of NMAE. To improve the accuracy of ARIMA, the hybrid models are introduced where the residuals of ARIMA are passed to both machine learning algorithms. ARIMA-RF and ARIMA-BCART models reveal a significant improvement in prediction. ARIMA-RF has reduced the error rate by 13% - 17% for different prediction period. ARIMA-BCART has slightly less accuracy than ARIMA-RF and improved the average accuracy by 10%-15%.

The comparison of the accuracy in terms of NMAE for the different datasets of the US East and West coasts with three significant weather variables are displayed in TABLE IV and Fig. 3. These results are similar to those depicted in TABLE III and Fig. 2; however, the accuracy has slightly increased for the three important variables. For different prediction intervals, ARIMA-RF has declined the error rate by 13% - 27%. ARIMA-BCART is marginally less robust than ARIMA-RF and has minimized the error rate by 10%-23% in an average.

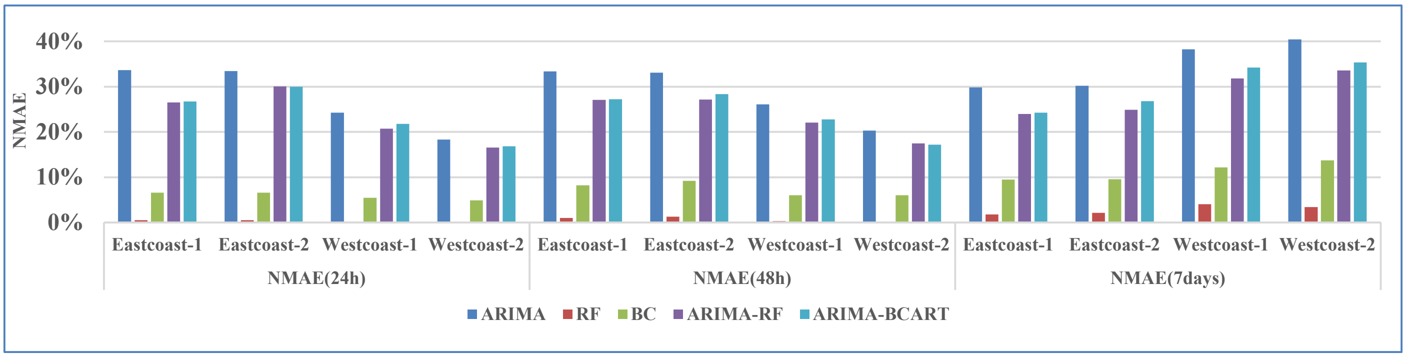

Fig. 2. Comparison of NMAE with different data sets containing five weather variables.

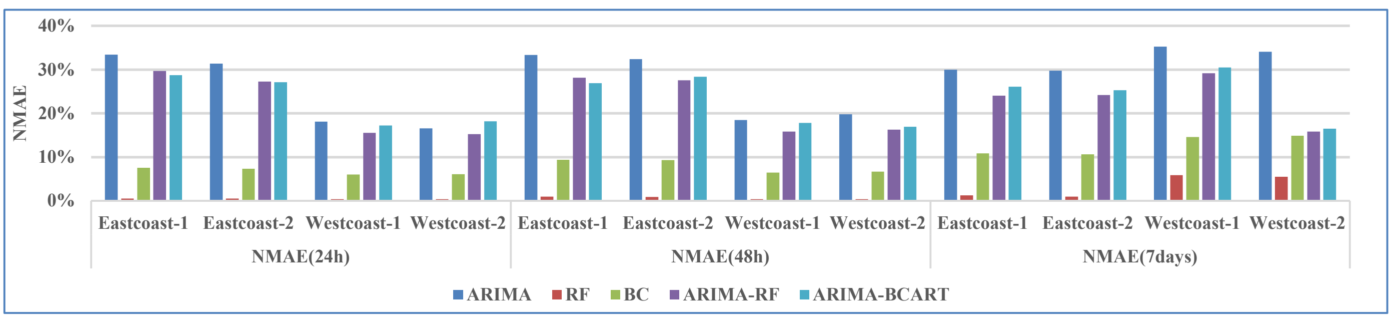

Fig. 3. Comparison of NMAE with different data sets containing three significant weather variables.

The comparison of the prediction accuracy in terms of NRMSE for the US East and West coast datasets with five weather variables are recorded in TABLE V and Fig. 4, and the three significant weather variables are shown in TABLE VI and Fig. 5.

TABLE V. COMPARISON OF NRMSE FOR FIVE WEATHER VARIABLE DATA

Time	Location	ARIMA	RF	BCART	ARIMA-RF	ARIMA-BCART
24H	Eastcoast-1	40.58%	0.61%	7.69%	31.17%	32.55%
	Eastcoast-2	41.37%	0.67%	7.07%	35.09%	35.70%
	Westcoast-1	28.26%	0.27%	6.58%	24.10%	25.33%
	Westcoast-2	38.62%	0.63%	6.60%	19.03%	18.79%
48H	Eastcoast-1	39.67%	2.63%	9.15%	32.74%	35.34%
	Eastcoast-2	39.83%	3.37%	11.02%	33.73%	34.43%
	Westcoast-1	29.13%	0.32%	6.80%	25.36%	25.55%
	Westcoast-2	37.18%	3.14%	10.29%	20.11%	19.94%
7 DAYS	Eastcoast-1	48.16%	4.95%	22.62%	38.15%	45.07%
	Eastcoast-2	48.59%	6.06%	22.68%	39.38%	45.44%
	Westcoast-1	64.43%	4.03%	31.78%	53.90%	60.72%
	Westcoast-2	67.13%	5.65%	21.18%	55.24%	59.25%

TABLE VI. COMPARISON OF NRMSE FOR THREE MOST IMPORTANT VARIABLES.

Time	Location	ARIMA	RF	BCART	ARIMA-RF	ARIMA-BCART
24H	Eastcoast-1	40.08%	0.60%	8.63%	34.48%	34.52%
	Eastcoast-2	36.96%	0.58%	8.59%	31.84%	32.13%
	Westcoast-1	21.19%	0.44%	7.20%	17.72%	20.05%

	Westcoast-2	23.38%	0.45%	7.29%	20.13%	20.26%
48H	Eastcoast-1	39.63%	2.21%	10.35%	32.88%	34.82%
	Eastcoast-2	37.87%	2.12%	10.35%	32.25%	34.22%
	Westcoast-1	21.14%	0.45%	7.40%	17.81%	19.39%
	Westcoast-2	22.07%	0.45%	7.55%	17.80%	18.48%
7 DAYS	Eastcoast-1	48.11%	3.58%	21.94%	38.02%	44.68%
	Eastcoast-2	49.23%	2.58%	21.90%	38.67%	44.88%
	Westcoast-1	65.90%	27.98%	39.06%	55.58%	57.98%
	Westcoast-2	63.99%	26.31%	39.07%	17.34%	17.69%

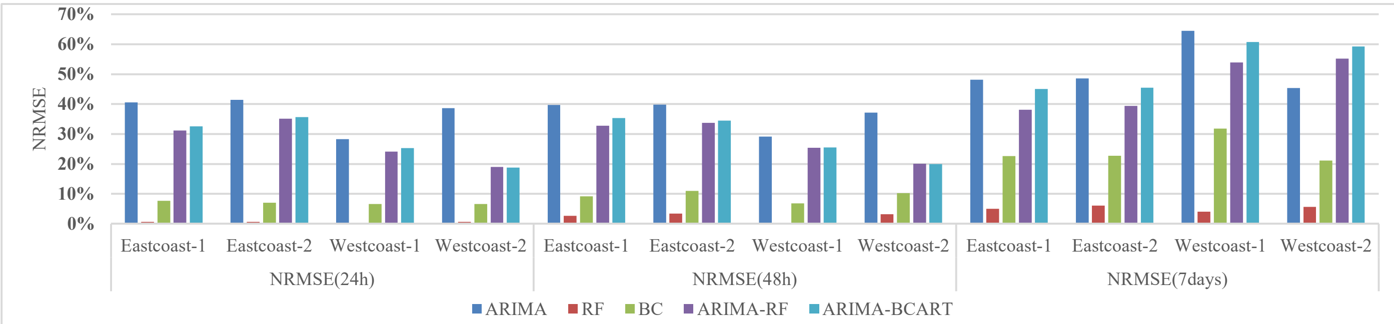

Fig. 4. Comparison of NRMSE of different datasets with five weather variables.

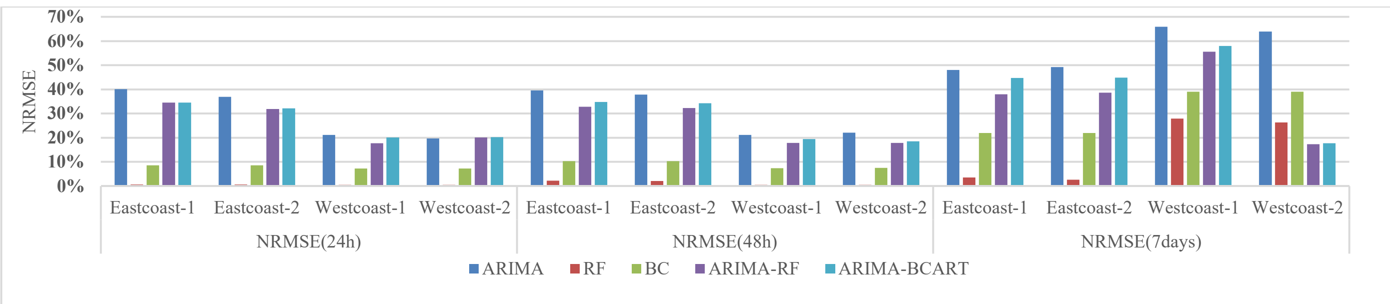

Fig. 5. Comparison of NRMSE of different datasets with three significant weather variables.

The results are as similar as the NMAE. Both ARIMA-RF and ARIMA-BCART have significantly improved the accuracy with five weather variables. ARIMA-RF has decreased the error rate by 18% - 26% for different prediction periods. ARIMA-BCART is marginally less powerful than ARIMA-RF and has increased the average accuracy by 8%-24%. Similar enhancements are noticeable for forecasting using three crucial weather parameters, and the ARIMA-RF and ARIMA-BCART have boosted the prediction accuracy by 15%-32% and by 12%-24% respectively.

VI. Conclusion

We have introduced prediction models for short-term and medium-term wind power generation for the US East and West coast data sets. The RF method yields the best performance for all datasets. The error rate is high for ARIMA; thus, the two-stage hybrid approaches have been proposed, and its implementation results are discussed. The hybrid models have improved the prediction accuracy of ARIMA. The results demonstrate that hybrid models are best suitable for time series data, where both linear and non-linear features are present. The performance of both ARIMA-RF and ARIMA-BCART models are similar for wind power prediction. The proposed hybrid models have boosted ARIMA's performance by 8% -26% in average. Future studies can be done on the response of hybrid methods to introducing random data into the critical parameters. High stochasticity of wind data requires a robust model for prediction, and analysis of this model's performance for highly arbitrary data will further demonstrate their effectiveness. Further studies can explore integration and evaluation of multi-stage forecasts models and ensemble methods into an optimization problem that reduces wind resource costs and dispatch planning scenarios by including storage constraints.

VII. Acknowledgement

This research work is sponsored by NSF award # 1537565 as part of EE 623-Introduction to Smart Grid Course at the University of North Dakota.

References

[1] C. Monteiro, R, Bessa, V., Miranda, A. Botterud, J. Wang, Jianhui, G. Conzelmann, G., Sciences, Decision, & Porto, INESC, Wind Power Forecasting: State-of-the-Art, Argonne National Laboratory, 2009.

[2] N. Sabrina, A. Sharif, R. Vishwajit, B. Stephen, Analysis and Application of Seasonal ARIMA Model in Energy Demand Forecasting: A Case Study of Small Scale Agricultural Load, Department of Electrical and Computer Engineering, Texas Tech University.

[3] U.S. Department of Energy (DOE), 20% Wind Energy by 2030: Increasing Wind Energy's Contribution to U.S. Electricity Supply, DOE Office of Energy Efficiency and Renewable Energy Report, July 2008. http://www1.eere.energy.gov/windandhydro/wind_2030.html.

[4] Yan, Y., Liu, S., Han, C. Gu, F. Li, A Robust Probabilistic Wind Power Forecasting Method Considering Wind Scenarios, 3rd Renewable Power Generation Conference (RPG 2014), Naples, pp. 1-6, 2014.

[5] R. Angamuthu Chinnathambi, A. Mukherjee, M. Campion, H. Salehfar, T.M. Hansen, J. Lin, P.A. Ranganathan, Multi-Stage Price Forecasting Model for Day-Ahead Electricity Markets, Forecasting 2018, vol. 1, pp. 26-46, [pdf]. http://www.mdpi.com/2571-9394/1/1/3, 2018.

[6] G. P. Zhang, Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model, Neurocomputing, Elsevier, vol. 50, pp. 159–175, 2003.

[7] R. Shumway and D. Stoffer, Time Series Analysis and Its Applications: With R Examples, ser. Springer Texts in Statistics, Springer New York, 2010.

[8] L. Breiman, Random Forests, Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[9] A. Mumtaz, P. Ramendra, X. Yong, Y. Zaher undher, Complete Ensemble Empirical Mode Decomposition Hybridized with Random Forest and Kernel Ridge Regression Model For Monthly Rainfall Forecasts, Journal of Hydrology, vol. 584, 2020.

[10] A. Callens, D. Morichon, S. Abadie, M. Delpey, L. Benoit, Using Random Forest and Gradient Boosting Trees to Improve Wave Forecast at a Specific Location, Applied Ocean Research, vol. 104, 2020.

[11] A. Lahouar and J. Ben Hadj Slama, Random Forests Model for One Day Ahead Load Forecasting, IREC2015 The Sixth International Renewable Energy Congress, Sousse, 2015.

[12] Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and Qsar Modeling, J. Chem. Inf. Comp. Sci., vol. 43, pp. 1947−1958, 2003.

[13] S. Basterrech and V. Snášel, Time-series Forecasting Using Bagging Techniques and Reservoir Computing, International Conference on Soft Computing and Pattern Recognition (SoCPaR), Hanoi, pp. 146-151, 2013.

[14] L. Breiman, Bagging Predictors, Department of Statistics, University of California, Berkley, California 94720, USA, Tech. Rep. 421, September 1994.

[15] S. Makridakis, A. Anderson, R. Carbone, R. Fildes, M. Hibdon, R. Lewandowski, J. Newton, E. Parzen, R. Winkler, The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition, J. Forecasting 1,111–153, 1982.

[16] J. G. De Gooijer and R. J. Hyndman, 25 Years of Time Series Forecasting, Int. J. Forecast., vol. 22, no. 3, pp. 443–473, 2006.

[17] S. Suhartono, S.P. Rahayu, D. Prastyo, D.G.P. Wijayanti, Juliyanto, Hybrid Model for Forecasting Time Series with Trend, Seasonal and Calendar Variation Patterns, Journal of Physics: Conference Series. 890. 012160. 10.1088/1742-6596/890/1/012160. (2017).

[18] National Renewable Energy Laboratory (NREL), Website: https://maps.nrel.gov/wind-prospector.