Applied Energy 190 (2017) 1245–1257

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

A data-driven multi-model methodology with deep feature selection for short-term wind forecasting

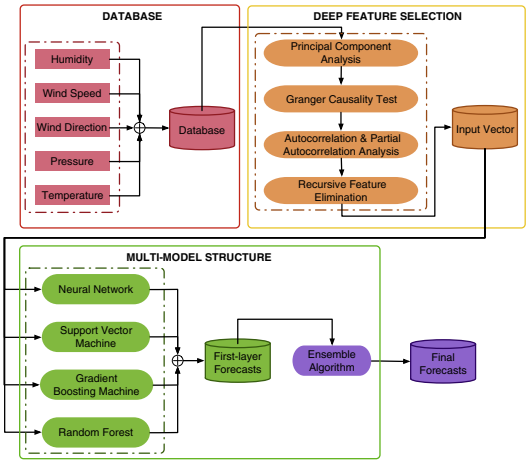
Cong Feng [a], Mingjian Cui [a], Bri-Mathias Hodge [b], Jie Zhang [a,*]

[a] Department of Mechanical Engineering, University of Texas at Dallas, Richardson, TX 75080, USA
[b] National Renewable Energy Laboratory, Golden, CO 80401, USA

HIGHLIGHTS

- An ensemble model is developed to produce both deterministic and probabilistic wind forecasts.
- A deep feature selection framework is developed to optimally determine the inputs to the forecasting methodology.
- The developed ensemble methodology has improved the forecasting accuracy by up to 30%.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 11 October 2016
Received in revised form 20 December 2016
Accepted 15 January 2017

Keywords:
Wind forecasting
Machine learning
Multi-model
Data-driven
Ensemble forecasting
Feature selection

ABSTRACT

With the growing wind penetration into the power system worldwide, improving wind power forecasting accuracy is becoming increasingly important to ensure continued economic and reliable power system operations. In this paper, a data-driven multi-model wind forecasting methodology is developed with a two-layer ensemble machine learning technique. The first layer is composed of multiple machine learning models that generate individual forecasts. A deep feature selection framework is developed to determine the most suitable inputs to the first layer machine learning models. Then, a blending algorithm is applied in the second layer to create an ensemble of the forecasts produced by first layer models and generate both deterministic and probabilistic forecasts. This two-layer model seeks to utilize the statistically different characteristics of each machine learning algorithm. A number of machine learning algorithms are selected and compared in both layers. This developed multi-model wind forecasting methodology is compared to several benchmarks. The effectiveness of the proposed methodology is evaluated to provide 1-hour-ahead wind speed forecasting at seven locations of the Surface Radiation network. Numerical results show that comparing to the single-algorithm models, the developed multi-model framework with deep feature selection procedure has improved the forecasting accuracy by up to 30%.

* Corresponding author.
E-mail address: jiezhang@utdallas.edu (J. Zhang).

http://dx.doi.org/10.1016/j.apenergy.2017.01.043
0306-2619/

1. Introduction

Renewable energy resources, particularly wind and solar energy, have become a primary focus in government policies, academic research, and the power industry. Among various renewables, wind energy is considered as one of the most promising alternatives [1]. However, the variable and uncertain nature of the wind resource may affect the economic and reliable operations of the power system [2], especially with the increasing penetration levels of wind power [3]. Therefore, it is important and desired to improve the accuracy of the wind speed forecasting (WSF) or wind power forecasting (WPF) that is used in power system scheduling. Different forecasting models have been developed in the literature, and they can be generally classified into three groups [4]: (i) physical models that are usually based on numerical weather prediction (NWP) models; (ii) statistical methods, most of which are intelligent algorithms based on data-driven approaches; and (iii) hybrid physical and statistical models.

NWP models simulate the physics of the atmosphere utilizing physical laws and boundary conditions. There exist a variety of challenges by directly adopting NWP models for wind forecasting, such as the accuracy, spatial and temporal resolutions, domain and hierarchical importance of the physical processes. Based on the domain coverage, the NWP models could be divided into limited area models (LAMs) and global models (GMs) [5]. Several GMs [6–8] have been developed to fulfill different forecasting needs, such as the Global Forecast System (GFS) and the Integrated Forecast Model (IFS). LAMs normally produce higher-resolution forecasts than GMs. Different LAMs have been developed for forecasting at different domains, some of which include the High-Resolution Limited Area Model (HIRLAM) [9], ALADIN [10], the Fifth-Generation Mesoscale Model (MM5) [11], and High Resolution Rapid Refresh (HRRR) [12].

Statistical models are trained using historical data and usually outperform NWP models in very short-term forecasting (within one-hour ahead) [13]. Both linear and non-linear methods have been widely applied to wind forecasting. Linear models, such as autoregressive moving average (ARMA) methods [14,15], Box-Jenkins methods [16], Kalman filter [17], and Markov Chain models [18,19], are most widely used in the literature. Artificial neural networks (ANN) and support vector machine (SVM) are the two most popular nonlinear methods for wind forecasting. Ghorbani et al. [20] forecasted one-hour ahead wind speed with ANN model combined with genetic expression programming. It was found that this model could significantly improve the forecast accuracy compared to the selected benchmark models. Chitsaz [21] adopted multi-dimensional wavelets as the activation functions in the ANN models to improve the forecasting accuracy. Li and Shi [22] comprehensively compared different ANN models in wind speed forecasting and concluded that the ANN models performed inconsistency with different conditions. Zhou et al. [23] developed three SVM methods with three kernels, and found that the SVM model performed better than the persistence approach for the test cases. Decision trees of many forms have also been used extensively as nonlinear methods for wind power forecasting. Troncoso et al. [24] proposed several regression tree models that could achieve competitive results with less computational time compared to other benchmark models. More research about wind forecasting with ANN and SVM algorithms has been done in [25–28].

Considering the spatial and temporal complexity of wind forecasting, it is challenging to develop a single algorithm that performs the best for all forecasting scenarios. An alternative way to reduce the risk of bad forecasts and improve the overall accuracy is hybriding multiple characteristically different algorithms, which is also known as 'ensemble forecasting'. Hybrid methods have been shown in the literature to produce more accurate forecasts than any of the individual forecasting models [29]. These hybrid or ensemble models can be divided into four categories: (i) data pre-processing based ensemble approaches; (ii) model-optimized ensemble approaches; (iii) data post-processing based ensemble approaches; and (iv) weighting-based ensemble approaches [30]. More information about the hybrid models could be found in [31–33]. For most studies, only two or three algorithms are blended with linear or non-linear weighting strategies. In addition, most models are tested with no more than two datasets, which is generally not enough to convincingly conclude that hybrid models are better than individual models, given the large differences in forecasting accuracy exhibited by the same algorithms at different sites. In this paper, a novel two-layer multi-model forecasting methodology is developed, which utilizes multiple characteristically different machine learning algorithms with different kernels in both layers. The developed methodology is validated with the data collected from seven Surface Radiation (SURFRAD) network locations to provide 1-h-ahead wind forecasting.

One of the major components of the developed multi-model forecasting model is a deep feature selection framework, which can select the most suitable inputs to the forecasting model. Full-dimension features will not only increase the computation time but also decrease the forecasting accuracy. A well-designed feature selection process plays a key role in wind forecasting. Different feature selection methods have been used in the literature. Liu et al. [34] utilized the autocorrelation function (ACF) and partial autocorrelation function (PACF) analysis, along with the Granger causality test to quantitatively analyze the relation between wind speeds and other variables on different lags. Kou et al. [35] used the sequential forward greedy search approach to determine the length of historical wind speed data as the inputs. Li et al. [36] developed a conditional mutual information-based feature selection approach to determine a small set of wind power and wind speed as input features. However, most of the existing feature selection methods present two major issues: (i) linear methods, such as ACF and PACF analysis, only consider the linear relations between time series; and (ii) nonlinear methods always take a continuous feature subset as a factor but do not analyze the individual lags.

To bridge the gap discussed above in wind forecasting, this paper develops a data-driven multi-model methodology with a deep feature selection process for short-term wind forecasting. The main contributions of this paper are as follows:

(1) An ensemble model for WSF is developed to automatically blend multiple single-algorithm models with different kernels being able to generate both deterministic and probabilistic forecasts.
(2) A deep feature selection framework is developed to optimally determine the input vector to the forecasting methodology.

The remainder of the paper is organized as follows. Section 2 describes the developed deep feature selection procedure and the individual algorithms employed in the two-layer hybrid model. Section 3 presents the results of deep feature selection, deterministic and probabilistic wind forecasts at multiple studied locations. Section 4 gives the conclusions and future work.

2. Multi-model wind forecasting with deep feature selection

Due to the nonlinear and non-stationary characteristics of wind speed, it is challenging to develop a generic model based on a single machine learning algorithm that can produce the best forecasts at different spatial and temporal scales. In this paper, a data-driven

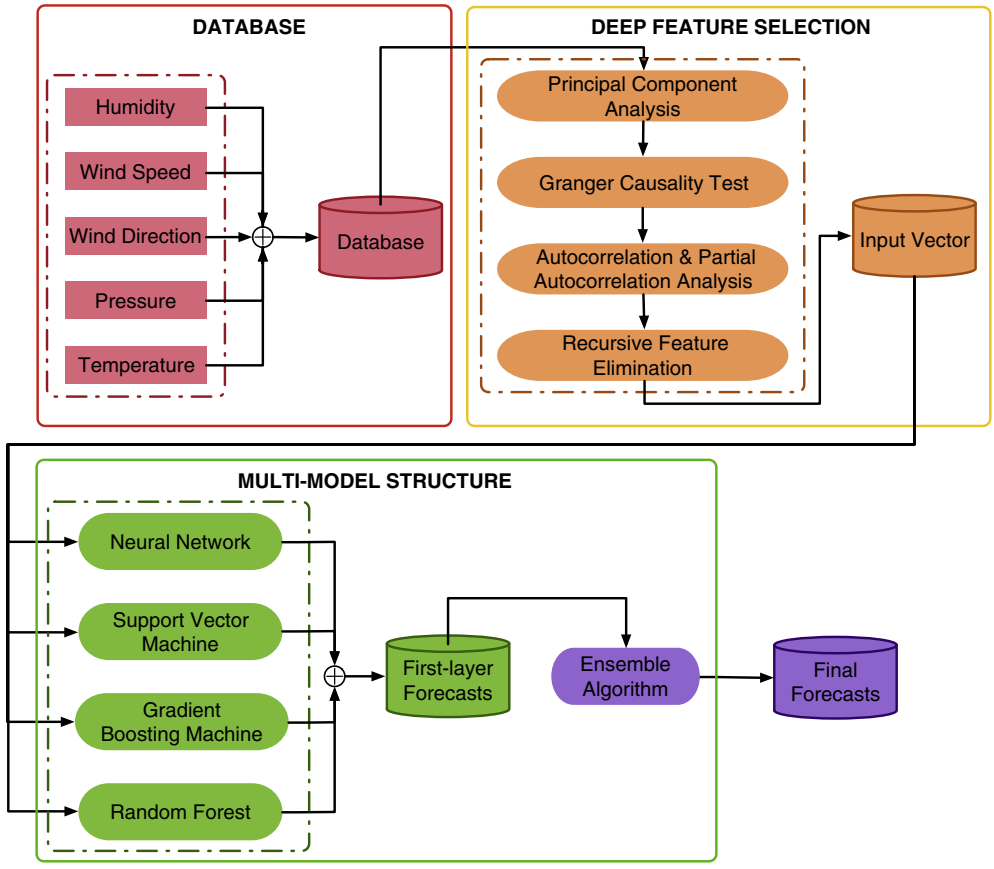


Fig. 1. The overall framework of the ensemble forecasting model.

multi-model wind forecasting methodology with deep feature selection is developed. The developed methodology is illustrated in Fig. 1, with a two-layer forecasting structure. First, features extracted from the data variables are determined by a deep feature selection procedure and serve as inputs to the model. Four independent feature selection methods are included in the procedure and implemented sequentially. The first layer machine learning models are built based on the selected feature combination. These models forecast wind speed or wind power as the output. A blending model is developed in the second layer to combine the forecasts produced by different algorithms from the first layer, and to generate both deterministic and probabilistic forecasts. Parameters of these models are optimally tuned by the grid search technique. Machine learning algorithms have distinctive advantages. For instance, ANN algorithms are adaptive by choosing different learning functions and loss functions, but have overfitting issues when the training data set is not long enough. SVM is efficient to train and can provide relatively accurate results, but they are memory-intensive and hard to tune. Tree ensemble algorithms like random forest and gradient boosting machine can avoid overfitting issues. The developed blending model is expected to integrate the advantages of different algorithms by canceling or smoothing the local forecasting errors.

2.1. Deep feature selection

The performance of a data-driven model highly depends on its inputs. There are several variables in one data set, such as wind speed, humidity, etc. Each variable has several lags, which are different features that need to be selected. The selected features will serve as inputs to the machine learning models. A comprehensive feature selection as illustrated in Fig. 2 is developed with the aim of improving the forecasting accuracy by selecting optimal feature combinations. Four different approaches are employed to select the most suitable input variables, which are: (i) principal component analysis (PCA); (ii) Granger causality test (GCT); (iii) autocorrelation analysis (ACF) and partial autocorrelation analysis (PACF); and (iv) recursive feature elimination (RFE).

2.1.1. Principal component analysis (PCA)

The data used for WPF or WSF usually contains a large number of variables, which may lead to high computational cost and may also decrease the prediction accuracy due to the extraneous information. To reduce the risk of over-fitting and inaccurate forecasts, PCA is applied to determine the major factors that contribute to the prediction [37]. The substance of PCA is the linear transformation. By using PCA, the observation matrix, X, is transformed into covariance matrix $\sum$. The contribution rate (CR) and the cumulative contribution (CC) of the ith principal component are, respectively, computed by:

$$CR_i = \frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j} \tag{1}$$

$$CC_i = \frac{\sum_{j=1}^{i}\lambda_i}{\sum_{j=1}^{p}\lambda_j} \tag{2}$$

where λ_i is the eigenvalue corresponding to the ith principal component, and p is the number of parameters.

2.1.2. Granger causality test (GCT)

PCA can reduce the variable dimension, but not all of the remaining variables are useful for the forecasting. To further explore correlations between the remaining variables and the wind

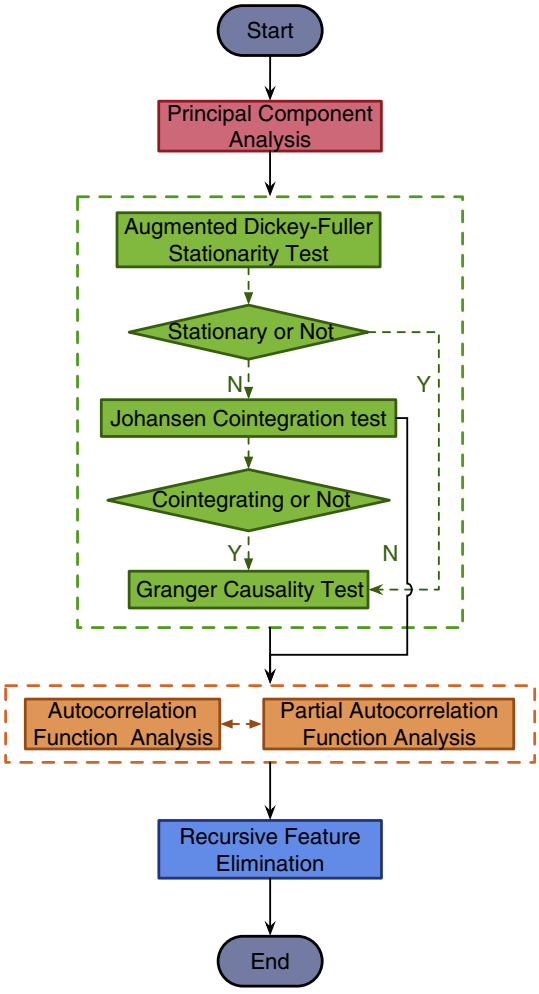


Fig. 2. The framework of the deep feature selection procedure.

speed series, GCT is conducted. GCT is a statistical hypothesis test first proposed by Clive W.J. Granger [38]. GCT has been widely applied in econometrics, and recently also in WPF or WSF. The efficiency of GCT in WPF or WSF feature selection has been proved in the literature [34,39,40].

The overall process of GCT is described in the green dash rectangular of Fig. 2. The prerequisite to conduct GCT is to ensure the testing time series be stationary [41]. To this end, the Augmented Dickey-Fuller unit root test (ADF) [42] is chosen as the stationary test approach in this paper. The null hypothesis of the ADF test is 'the testing time series does not have unit root', which means the series is non-stationary. If the test statistic is smaller than the critical value, the proposed null hypothesis is rejected, which means the series is stationary and thus GCT can not be implemented directly. To further ensure the long-term relationship between two variables, the Johansen cointegration test (JCT) is also applied.

If the testing series is stationary or is validated to have stable relationship, GCT can be carried out. The unrestricted model and restricted model for the testing are, respectively, described as [38]:

$$x_1^u(t) = \sum_{j=1}^{l} \alpha_j x_1(t-j) + \sum_{j=1}^{p} \beta_j x_2(t-j) + \varepsilon(t) \tag{3}$$

$$x_1^r(t) = \sum_{j=1}^{l} \alpha_j x_1(t-j) + \varepsilon(t) \tag{4}$$

where $\{x_1\}$ and $\{x_2\}$ are testing variables, $\varepsilon(t)$ is the residual for the model, l and p are lags of series $\{x_1\}$ and $\{x_2\}$, respectively. The difference between the unrestricted model $x_1^u(t)$ and restricted model x_1^r is that: $x_1^u(t)$ contains the causal series $\{x_2\}$, while $x_1^r(t)$ does not contain the causal series $\{x_2\}$.

To determine whether $\{x_2\}$ Granger causes $\{x_1\}$ or not, the F-test is conducted as:

$$F = \frac{SSR_r - SSR_u}{SSR_u}\left(\frac{n-l-q}{q}\right) \tag{5}$$

where SSR_r is the sum of squared residuals of the restricted model, SSR_u is the sum of squared residuals of the unrestricted model, n is the sample size, q is the number of variables in Eq. (3), and $(l+q)$ is the number of variables in Eq. (4). The F-statistic is then compared to the critical value and the decision can be made based on the probability.

2.1.3. Autocorrelation and partial autocorrelation analysis

Both PCA and GCT methods identify the most important variables for the machine learning models, but don't consider different lags of each variable. In time series analysis, the autocorrelation function (ACF) and partial autocorrelation function (PACF) are two popular approaches to measure how a variable series is correlated with itself at different lags. Using ACF and PACF analysis, the most meaningful lags can be identified for forecasting. ACF indicates the correlation of the variables between two lags ρ_h, which is defined as:

$$\rho_h = Corr(x_{1t}, x_{1(t-h)}) = \frac{\gamma_h}{\gamma_0} \tag{6}$$

where x_{1t} is the wind speed at time t, $x_{1(t-h)}$ is the wind speed at time $(t-h)$, $\mathrm{r_h}$ is the covariance of wind speed lag h, and γ_0 is the covariance of current wind speed.

PACF denotes the correlation between variables at lag h and lag $(t-h)$ by removing all the dependence on other variables between the two lags, which is defined as:

$$\phi_h = Corr\{[x_t - P(x_t|x_{t-h+1},\ldots,x_{t-1})],[x_{t-h} - P(x_t|x_{t-h+1},\ldots,x_{t-1})]\} \tag{7}$$

where $P(A|B)$ is the correlation between A and B.

Then the confidence intervals are used to judge the significance of the autocorrelations between lags. One of the most widely used definition of the 95% confidence interval is defined by:

$$r_{.95} = \pm\frac{2}{\sqrt{N}} \tag{8}$$

where N is the data size. In this paper, both ACF and PACF are employed to determine the optimal wind speed lags for the machine learning inputs.

2.1.4. Recursive Feature Elimination (RFE)

From the previous feature selection procedure, redundant variables are filtered out and the remaining variables are $\{X_1, X_2, \ldots, X_m\}$. Additionally, the wind speed series has been analyzed and its most useful lags have been determined. However, each wind speed lag is evaluated separately, which may not guarantee the effectiveness of their combination with other variables' lags (e.g., temperature, pressure, etc.). Therefore, the RFE method is adopted to find the optimal combination of different lags of the filtered variables.

RFE is a type of wrapper method. It first trains the model with the original feature set $\{x_{i,1}, x_{i,2}, \ldots, x_{i,p}\}(i = 1, 2 \ldots, m)$ and ranks the features based on their importance. Then the model performance is evaluated based on different metrics. This procedure is repeated with a progressively smaller subset, which is reduced

by d features. The model with the best performance is picked out and the feature combination is determined. The algorithm is illustrated by a pseudo-code with the random forest algorithm in Fig. 3. The root mean square error (RMSE) metric is used to evaluate the model performance. Then the random forest algorithm is used to train the model. The RFE results are then compared and combined with the previous results.

2.2. Blending models and machine learning algorithms

With the suitable inputs selected by the deep feature selection procedure, the single-algorithm machine learning models can be expressed as follows,

$$y_i = f_i(x_1, x_2, \ldots, x_p) \tag{9}$$

where $f_i(*)$ is the ith algorithm and y_i is the wind speed forecasted by $f_i(*)$. The final forecasted wind speed by the blending algorithm, $\hat{y}$, is represented by:

$$\hat{y} = \Phi(y_1, y_2, \ldots, y_m) \tag{10}$$

To obtain accurate forecasts, multiple machine learning models from the first layer are included in the ensembles. In the second layer, several blending algorithms are also tested for generalizability. The algorithms employed in both two layers include artificial neural networks (ANN), support vector regression (SVR), gradient boosting machine (GBM), and random forest (RF) regression. The forecasting accuracy and computational cost are the two major criteria for selecting the different algorithms. All of these component algorithms used to develop the framework are selected from the state-of-the-art machine learning algorithms. These algorithms have been shown to perform well at different forecasting horizons in the literature [43–46]. The selected models have shown a similar level of performance and acceptable computational cost in the literature. Other algorithms like linear regression and time series models may add noise to the framework, and more advanced models like deep learning algorithm models require more computational power. Thus, they are not considered in the developed framework. The four selected component machine learning algorithms are briefly introduced in the following paragraphs.

ANN is a widely used algorithm that consists of interconnected neurons. ANN can be classified into different types with different activation functions and learning algorithms. The mathematical description of the ANN is expressed as:

$$y_i^{(n)} = f\left(\sum_{j=1}^{N} w_{ij}^{(n,n-1)} y_j^{(n-1)} + \theta_i^n\right) \tag{11}$$

where i is a neuron of the nth layer, w_{ij} is the weight from the neuron j in the layer $(n-1)$ to the neuron i in layer n, and θ_i^n is the threshold of the neuron i in the n-th layer.

SVM is a linear classifier proposed by Vapnik [47]. When dealing with linearly inseparable data, nonlinear mapping based kernel methods, $\kappa(x): \mathbb{R}^n \to \mathbb{R}^{n_h}$, are used to map the nonlinear data into the high dimensional feature space. Then, a linear hyper plane is found by maximizing the distance between support vectors and the hyper plane. The SVM algorithm can also be applied in regression problems, which is called support vector regression (SVR). The hyper plane function, also called the SVR function, is described as:

$$f(x) = \omega^T \kappa(x) + b \tag{12}$$

where ω and b are variables solved by minimizing the empirical risk, which is given by:

$$R(f) = \frac{1}{n}\sum_{i=1}^{n} \Theta_\varepsilon(y_i, f(x)) \tag{13}$$

where $\Theta_\varepsilon(y,f)$ is the ε-insensitive loss function, expressed as:

$$\Theta_\varepsilon(y,f) = \begin{cases} \|f-y\| - \varepsilon, & if\ \|f-y\| \geqslant \varepsilon \\ 0, & otherwise \end{cases} \tag{14}$$

Then the optimal hyper plane is found by solving the inequality-constrained quadratic optimization problem.

GBM is a type of ensemble learning method that implements the sequential boosting algorithm. Basically, the objective of GBM is to minimize the expectation of the loss function [33]. To achieve this, the residual of the initial model is calculated. Then the base learner is fitted to the residual by the gradient descent algorithm. Thus, the model is updated by adding the weighted base learner to the previous model. Finally, the target model is obtained by iteratively conducting the previous steps. The GBM algorithm is illustrated by the pseudo-code [48] in Fig. 4.

Random forest (RF) regression is another ensemble learning method that consists of many single classification and regression trees (CART). To train these single CART, the bagging algorithm is used to create different bootstrap samples from the input data. During this process, one third of the data are not contained in the bootstrap sample, which are called out-of-bag (OOB) data. These OBB data are used to test the CART tree. With all the single CART grown, the final prediction is made from aggregating the CART. Since RF is a combination of various different regressions, the model is generally free from over-fitting [49].

3. Case studies

3.1. Data source and pre-analysis

The developed multi-model wind forecasting methodology is applied to the data collected from the Surface Radiation Network (SURFRAD), which includes seven stations with diverse climates. Both deterministic and probabilistic wind forecasts are generated at the 1-h-ahead timescale. The information of the locations is briefly summarized in Table 1. The SURFRAD data contain more than twenty meteorological parameters, five of which are used in the research, including temperature, humidity, wind speed, wind

	Algorithm 1: Recursive Feature Elimination (RFE)
1	Train the random forest model with full feature set
2	Evaluate the model performance with RMSE and rank feature importance
3	**For** *i=1 to n*, **do**:
4	Eliminate last *d* features with smallest importance
5	Train the random forest model with tuned subset
6	Evaluate the model performance with RMSE and rank feature importance
7	**End** For
8	Select the optimal feature length and it's feature rank
9	**End** Algorithm

Fig. 3. RFE algorithm.

	Algorithm 2: Gradient Boosting Machine (GBM)
1	$F_0(x) = \arg\min_\rho \sum_{i=1}^{N} \Psi(y_i, \rho)$
2	**For** m=1 to M **do**:
3	$\bar{y}_i = -\left[\frac{\partial \Psi(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, N$
4	$a_m = \arg min_{a,\beta} \sum_{i=1}^{N} [\bar{y}_i - \beta h(x_i, a)]^2$
5	$\rho_m = \arg min_\rho \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(x_i) + \rho h(x_i, a_m))$
6	$F_m(x) = F_{m-1}(x) + \rho_m h(x, a_m)$
7	**End** For
8	**End** Algorithm

Fig. 4. Gradient boosting machine algorithm.

direction, and pressure. These meteorological variables have been used as the inputs to short-term wind forecasting models in the literature, such as [50,51]. Even though the wind speeds were recorded at the height of 10 meters rather than the hub height (more than 50 meters) of large wind turbines, the forecasting results can be used in a number of aspects, including but not limited to:

- Be directly used by small-scale or distributed wind turbines;
- Be scaled up and utilized by large-scale wind turbines;
- Provide weather forecasting results;
- Serve as inputs of solar panel power generation model for solar forecasting.

The data at each location are recorded every minute. The data period used in the research is from 2015-01-01 to 2015-12-31. The data summary and the pre-processing results are also listed in Table 1. A quality filter is applied to perform data pre-processing, which serves two main roles: (i) noisy data pre-processing by correcting abnormal data, and (ii) data trimming by averaging the minute data into hourly data. The scatter plots of the data smoothed by the quality filter are shown as a matrix in Fig. 5. Most of the data points are concentrated at a certain area, which are shown as green ellipses. The non-linear relationships between each pair of variables are also depicted in Fig. 5.

The first 1/3 of data are assigned as the training data for the first-layer models. Then, the first layer models forecast the 1-h-ahead wind speed with the second 1/3 of the data. Forecasts from the first-layer models together with the actual wind speed are used to train the second-layer model. The effectiveness of the developed multi-model framework is validated by the last 1/3 of data.

3.2. Deep feature selection case study

The deep feature selection procedure is applied to the data of seven locations. The results at the BND station are discussed in detail.

3.2.1. PCA feature selection

The five variables in the dataset are first processed with the PCA to determine the minimum necessary number of variables. The contribution rates of the principal components are listed in Table 2. It is seen from the table that the wind direction only contributes 5.7% in the data. The major information is retained even though the wind direction is left out. Thus the first four variables, wind speed, temperature, humidity, and pressure, are selected as inputs to the machine learning models.

3.2.2. GCT feature selection

With the remaining four variables, GCT is applied to determine the causality between each variable and the wind speed. Before implementing the GCT, the stationarity of each variable is checked by the ADF unit root test. If the variable series is not stationary, GCT cannot be applied directly. For the non-stationary variables, JCT is conducted to ensure the long-term relationship between the two variables. If there is a long-term relationship between the test variable and the wind speed, GCT can be carried out. If the test variable is non-stationary and doesn't have a long-term relationship with the wind speed, GCT cannot be applied. The test results are summarized in Table 3. It is observed that all of the test statistics exceed the pre-determined critical value for 99% confidence level. Hence, we are not able to reject the null hypotheses, indicating that these four variable series are non-stationary. Thus, GCT can not be implemented directly to the dataset.

To further confirm the applicability of the GCT, JCT is applied to test the long-term relationships between non-stationary series. The null hypothesis of $R = 0$ and $R \leq 1$ is 'there is no cointegration equation or there is at most one cointegration equation between the testing variable and the wind speed'. On the contrary to ADF test, if the test statistic is larger than the critical value, the null hypothesis is rejected and vice versa. As shown in Table 4, the test statistics of "$R = 0$ hypothesis" exceed the critical value for all of the three testing groups. It means there is at least one cointegration relationship between the each testing variable and the wind speed. And for the hypothesis of $R \leq 1$, all the test statistics are less than the critical value. It is implied that each testing variable has at most one cointegration relationship with the wind speed. Both the trace test and the maximum eigenvalue test indicate consistent results. Thus, there exist long-term relationships between wind speed and three other variables. Hence, the GCT test can be conducted.

Based on Eqs. (3)–(5), the GCT results are calculated. For each test variable, it is used as the input and the output of the forecasting models separately. If the variable is found to Granger-cause wind speed, this variable is valuable for wind speed prediction. If the wind speed is the Granger-cause of the test variable, it means the test variable is not useful to predict the wind speed. Table 5 illustrates the probability to reject the null hypothesis of GCT. The assumptions that 'Temperature does not Granger-cause wind speed' and 'Humidity does not Granger-cause wind speed' are rejected under 0.005 confident level. The causality between the

Table 1
SURFRAD stations summary.

Locations	State	Lat.	Long.	Elev. (m)	Sample no.	Bad data percentage (%)	Wind Speed	
							Mean (m/s)	SD (m/s)
BND	IL	40.05	−88.37	230	524,754	0.24	4.86	2.78
TBL	CO	40.12	−105.24	1689	525,092	1.81	3.09	2.17
DRA	NV	36.62	−116.02	1007	525,510	1.11	3.70	2.52
FPK	MT	48.31	−105.24	98	525,020	0.39	4.28	2.88
GCM	MS	34.25	−89.87	98	523,898	2.22	1.92	1.32
PSU	PA	40.72	−77.93	375	524,927	1.82	2.89	2.15
SXF	SD	43.73	−96.62	473	525,402	0.22	4.16	2.16

Note: More information about the SURFRAD locations can be found at the SURFRAD website (http://www.esrl.noaa.gov/gmd/grad/surfrad/).

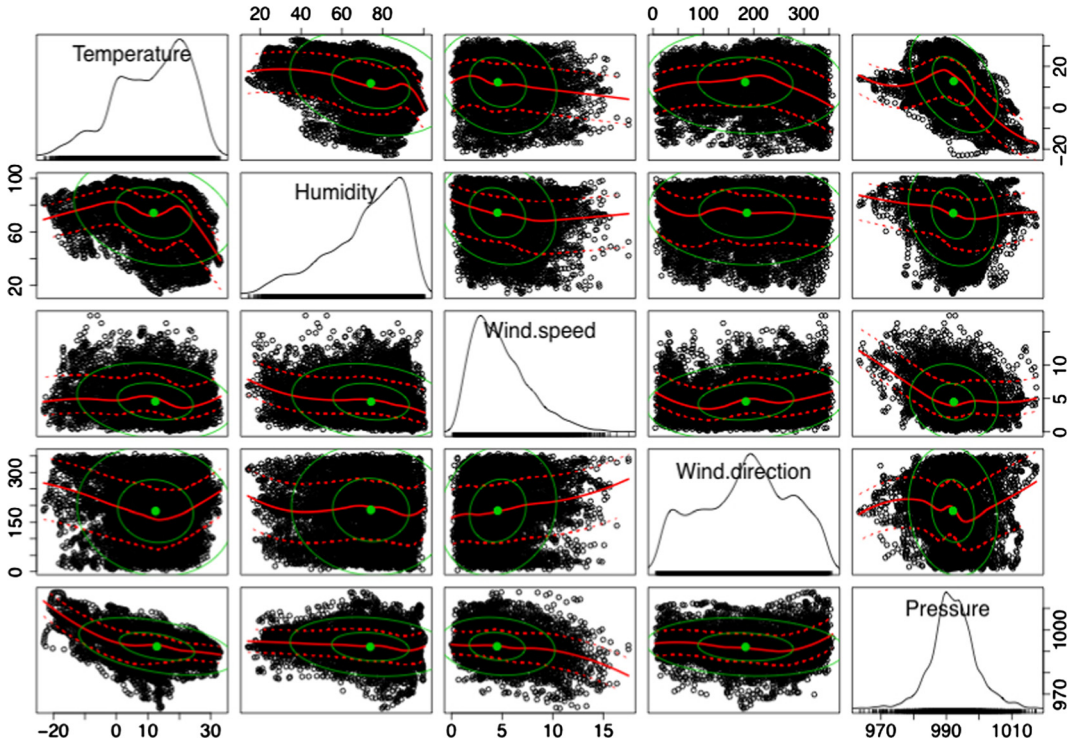


Fig. 5. Variable scatter plot matrix of BND data. Note that five variables are contained in the scatter plot: wind speed (m/s), temperature (deg. C), humidity (%), pressure (100 Pa), and wind direction (deg).

Table 2
Contribution rates of principal components.

Contribution rate (%)					Cumulative contribution (%)
Wind speed	Temperature	Humidity	Pressure	Wind direction	
29.1					29.1
29.1	24.9				54.0
29.1	24.9	22.7			76.7
29.1	24.9	22.7	17.6		**94.3**
29.1	24.9	22.7	17.6	5.7	100

Note: The contribution rate and cumulative contribution are calculated based on Eqs. (1) and (2), respectively. The number in boldface is the cumulative contribution of the retaining variables.

Table 3
ADF unit root test on BND data.

	Test variables			
	Wind speed	Temperature	Humidity	Pressure
Test statistic	0.2944	−0.5779	−0.2417	−0.1326
Critical value	−3.43	−3.43	−3.43	−3.43
Conclusion	N	N	N	N

Note: The critical value used in the research is for 99% confidence level. Y means stationary and N means non-stationary.

Table 4
Johansen cointegration test on BND data.

No. of CE(s)	Test statistic					
	Temperature		Humidity		Pressure	
	Trace	Maximum eigenvalue	Trace	Maximum eigenvalue	Trace	Maximum eigenvalue
$R = 0$	22.53 (19.96)	16.74 (15.67)	20.66 (19.96)	16.79 (15.67)	27.85 (19.96)	20.70 (15.67)
$R \leq 1$	5.79 (9.24)	5.79 (9.24)	7.08 (9.24)	8.24 (9.24)	7.15 (9.24)	7.15 (9.24)

Note: The critical values of 95% confidence level are shown in parentheses. CE means cointegration equation.

Table 5
Granger causality test results.

Test variables	Hypotheses	Prob.
Temperature	Temperature does not Granger-cause wind speed	0.0002
	Wind speed does not Granger-cause temperature	0.1067
Humidity	Humidity does not Granger-cause wind speed	0.0007
	Wind speed does not Granger-cause humidity	0.0346
Pressure	Pressure does not Granger-cause wind speed	0.7567
	Wind speed does not Granger-cause pressure	0.0277

pressure and the wind speed is not significant. Thus, historical wind speed, temperature, and humidity are chosen as the useful inputs to the wind speed prediction.

3.2.3. ACF and PACF feature selection

In WSF, the most important features are wind speeds of previous hours. ACF and PACF analyses are employed to determine the dependence of wind speed lags. The correlograms and partial correlograms are plotted in Fig. 6. It is shown that the wind speed series has significant autocorrelation between lag 0 (itself), lag 1, and lag 2. After removing the internal relations between each two lags, the partial correlation is obtained, which is illustrated by partial correlograms in Fig. 6. The partial correlation is significant between lag 0 and lag 1. Even though the partial correlation coefficient does not exceed the 95% confident interval (the blue dash lines), lag 2 is still considered as an important feature for the forecasting models. This is because its value is not negligible compared to other lags, which is consistent with the results from ACF analysis.

3.2.4. RFE feature selection

The relationship between the lags of other two variables (temperature and humidity) and the objective wind speed also needs to be explored. Therefore, to ensure the performance of the selected feature combination, RFE is used as the last step in the deep feature selection procedure. Different numbers of features are applied as inputs to the RF models and the performance is evaluated in terms of RMSE. Fig. 7 shows the results of model evaluations with different input combinations. The models with too few inputs show an unsatisfactory accuracy, while the models with too many inputs are computationally prohibitive. The best model has an RMSE of 0.897 with 30 features as inputs, as highlighted by the triangle in Fig. 7. Compared to the model only using the current wind speed as the input, the accuracy is improved by up to 12.4%. However,

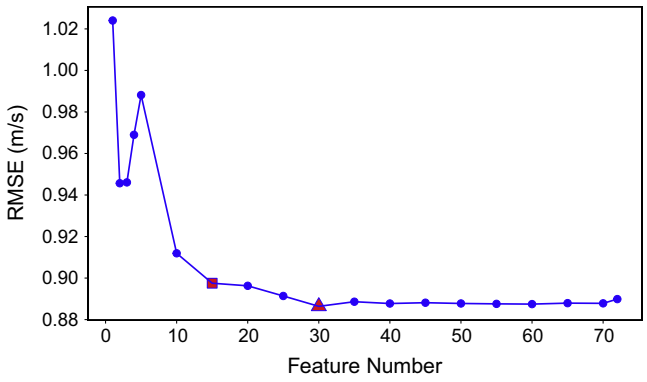


Fig. 7. Model performance across different subset sizes.

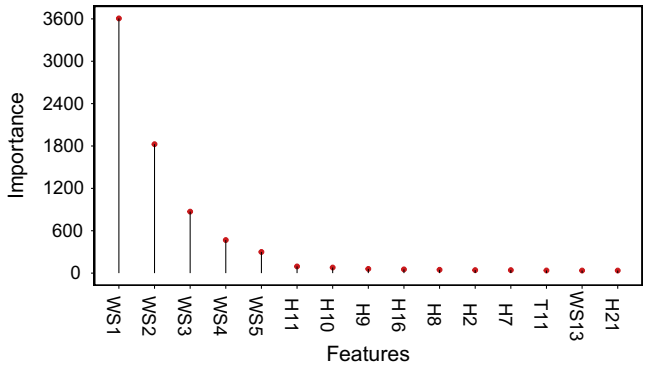


Fig. 8. Importance ranks of the selected 15 features.

30-feature inputs are too computationally expensive for training the model. The forecasting model with 15-feature inputs also generates competitively accurate results however with much less computation time (235.2 h for one location, using a workstation with 1.6 GHz processor and 36 GB RAM), shown by the rectangle in Fig. 7. Thus, the 15 features subset is finally chosen as the inputs to wind speed forecasting at BND. The selected features and their importance ranks are shown in Fig. 8. The first five features *WS*1, *WS*2, *WS*3, *WS*4, and *WS*5 are much more important than the other ten features as inputs to the forecasting model. However, if only the first five features are employed as forecasting inputs, the

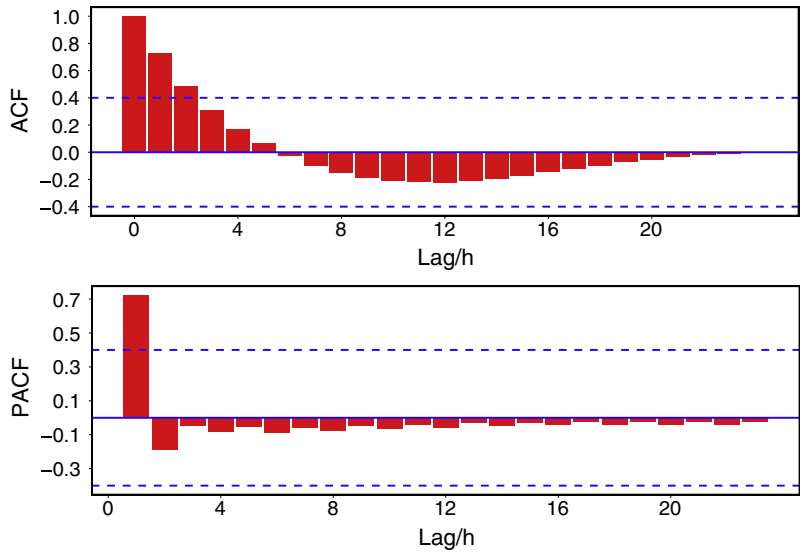


Fig. 6. Correlogram and partial correlogram of wind speed lags at BND.

Table 6
Results of parameter tuning at the BND site.

Model	Search range	Selected value
SVR_li	$C \in [1:100]$	$C = 1$
SVR_poly	$C \in [1:100]; degree \in [1:5]$	$C = 3; degree = 3$
ANN	$n_l \in [1:5]; n_o \in [1:50]; decay \in [0.01:1]$	$n_l = 1; n_o = 36; decay = 0.04$
GBM	$n.trees \in [50:1500]; \lambda \in 0.1^{[1:4]}; int.depth \in [1:10]; n.ob \in 5^{[1:10]}$	$n.trees = 650; \lambda = 0.01; int.depth = 9; n.ob = 5$
RF	$mtry \in [1:100]$	$mtry = 7$

performance of the forecasting model is much worse than that with all fifteen features combined, as illustrated in Fig. 7.

3.3. Model selection and parameter tuning

The developed multi-model framework includes multiple individual models in the first layer and also several models in the second layer. Different algorithms are tested in both layers. The parameters are optimized to improve the forecasting performance of these single-algorithm models. In this paper, the grid search

Table 7
1-h-ahead forecasting NMAE of single-algorithm models without feature selection.

Models	BND	TBL	DRA	FPK	GCM	PSU	SXF
P	**4.05**	**4.27**	**5.25**	**4.28**	**4.13**	**5.78**	**3.91**
SVR_li	5.26	5.04	6.65	5.18	5.42	7.13	4.93
SVR_poly	5.04	4.90	6.17	4.93	5.06	6.86	4.86
ANN	5.35	5.96	6.23	5.29	5.65	6.90	4.73
GBM_g	4.95	4.82	6.02	4.80	4.82	6.68	4.78
GBM_l	5.01	4.80	6.23	4.94	4.96	6.67	4.93
RF	5.32	4.93	6.51	5.31	5.58	7.51	5.25

Note: The best NMAE values of the component models are in boldface. P represents the persistence model.

Table 8
1-h-ahead forecasting NRMSE of single-algorithm models without feature selection.

Models	BND	TBL	DRA	FPK	GCM	PSU	SXF
P	**5.65**	**6.60**	**7.36**	**5.91**	**5.68**	**8.27**	**5.42**
SVR_li	7.76	8.37	9.88	7.92	8.09	9.90	6.95
SVR_poly	7.05	7.62	8.58	6.81	6.72	9.33	6.51
ANN	7.27	8.09	8.47	6.94	7.05	9.37	6.30
GBM_g	6.78	7.77	8.06	6.59	7.01	9.24	6.37
GBM_l	6.79	7.71	8.86	6.68	6.67	9.42	6.52
RF	7.36	7.21	9.10	7.35	7.46	10.04	7.11

Note: The best NRMSE values of the component models are in boldface. P represents the persistence model.

Table 9
NMAE of the developed multi-model forecasting without feature selection.

Models	BND	TBL	DRA	FPK	GCM	PSU	SXF
P	**4.05**	**4.27**	**5.25**	**4.28**	**4.13**	**5.78**	**3.91**
E_SVM_li	4.32	5.28	5.44	4.45	6.04	6.03	4.05
E_SVM_poly	*4.20*	*4.54*	*5.36*	*4.31*	*5.14*	*5.84*	*4.01*
E_GBM	4.26	4.58	5.49	4.37	5.81	6.11	4.19
E_RF	4.26	4.60	5.66	4.33	5.34	6.09	4.22

Note: The smallest NMAE at each location is in boldface. The best ensemble model is in italic.

Table 10
NRMSE of the developed multi-model forecasting without feature selection.

Models	BND	TBL	DRA	FPK	GCM	PSU	SXF
P	**5.65**	**6.60**	**7.36**	**5.91**	**5.68**	**8.27**	**5.42**
E_SVM_li	6.20	8.96	7.51	6.29	9.21	8.52	5.61
E_SVM_poly	*5.77*	*7.22*	*7.36*	6.05	*7.08*	*8.16*	*5.49*
E_GBM	5.95	7.29	7.58	6.00	8.23	8.48	5.72
E_RF	5.85	7.52	7.63	*5.92*	7.53	8.46	5.74

Note: The smallest NRMSE at each location is in boldface. The best ensemble model is in italic.

Table 11
NMAE of the developed multi-model forecasting with deep feature selection.

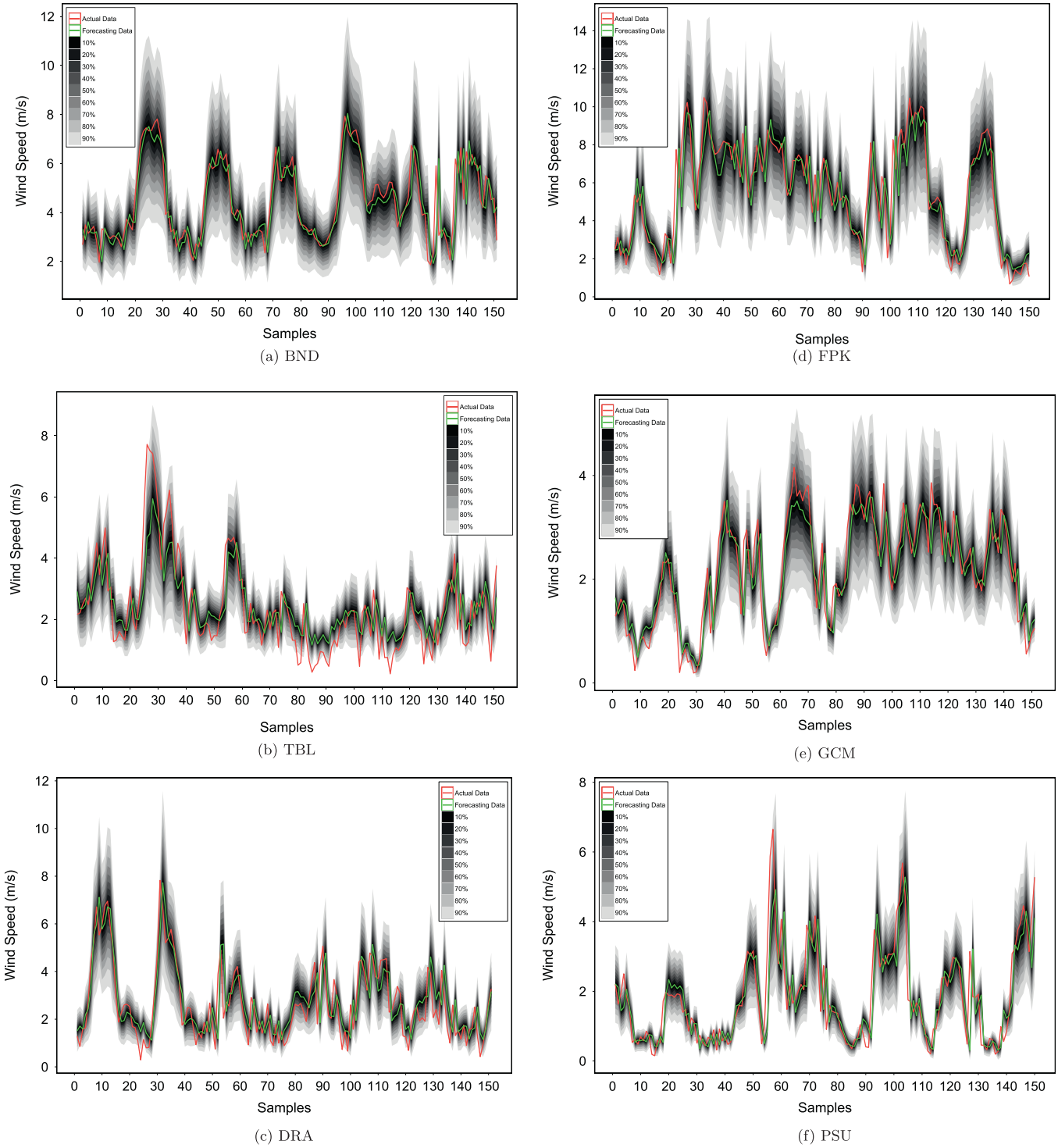
Models	BND	TBL	DRA	FPK	GCM	PSU	SXF
P	4.05	4.27	5.25	4.28	**4.13**	5.78	3.91
E_SVM_li	4.05	4.61	*5.20*	*4.19*	4.76	*5.72*	*3.84*
E_SVM_poly	**3.93**	**4.12**	*5.21*	**4.08**	**4.13**	**5.70**	**3.76**
E_GBM	4.13	4.40	**5.18**	*4.20*	4.54	*5.77*	*3.82*
E_RF	4.21	4.51	5.30	*4.22*	4.54	5.86	*3.87*

Note: The smallest NMAE at each location is in boldface. Other NMAEs smaller than persistence model are in italic.

Table 12
NRMSE of the developed multi-model forecasting with deep feature selection.

Models	BND	TBL	DRA	FPK	GCM	PSU	SXF
P	5.65	6.60	7.36	5.91	5.68	8.27	5.42
E_SVM_li	5.69	8.12	7.36	*5.82*	7.91	*8.10*	*5.35*
E_SVM_poly	**5.43**	**6.17**	**7.24**	**5.64**	**5.67**	**8.09**	**5.21**
E_GBM	5.83	6.93	**7.24**	*5.81*	6.65	*8.11*	*5.31*
E_RF	5.85	7.02	*7.30*	*5.77*	6.47	*8.20*	*5.34*

Note: The smallest NRMSE at each location is in boldface. Other NRMSEs smaller than persistence model are in italic.

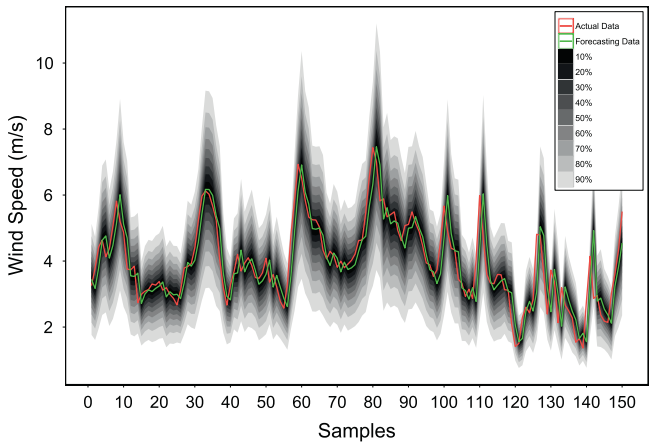


(a) BND (d) FPK

(b) TBL (e) GCM

(c) DRA (f) PSU

Fig. 9. Deterministic forecasting from the multi-model framework with confidence intervals



(g) SXF

Fig. 9 (*continued*)

method is used to determine the optimal parameters that generate the minimum forecasting error.

For SVR, the linear (SVR_li) and polynomial (SVR_poly) models are selected in the first layer. The sole parameter of SVR_li to tune is the cost (C). For SVR_poly, there are two parameters to be determined: polynomial degree (*degree*) and cost (C).

For ANN, different learning algorithms and activation functions are tested. The selected models employ the feed-forward back propagation as the learning function, and the sigmoid function as the activation function. The most important parameters for ANN models are the hidden layer number (n_l), neurons in each layer (n_o), and weight decay parameter (*decay*).

For the GBM models, different loss functions are utilized. Two GBM models are selected in the first layer using Gaussian (GBM_g) and Laplacian (GBM_l) loss functions. Four parameters, i.e., the number of trees to fit (*n.trees*), the learning rate (λ), the maximum depth of variable interactions (*int.depth*), and the minimum number of observations in the terminal nodes (*n.ob*) need to be tuned.

The RF model in this paper only has one parameter to be optimized. It is the number of variables randomly sampled as candidates at each split (*mtry*). One example of parameters optimization results is listed in Table 6.

3.4. Deterministic results of the multi-model forecasting

In order to evaluate the forecasting accuracy of the developed framework, two error criteria are utilized: the normalized mean absolute error (*NMAE*) and the normalized root mean square error (*NRMSE*). They are defined by:

$$NMAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f_i - y_i}{y_{max}}\right| \tag{15}$$

$$NRMSE = \frac{1}{y_{max}}\sqrt{\frac{\sum_{i=1}^{n}(f_i - y_i)^2}{n}} \tag{16}$$

where f_i is the forecasted wind speed, y_i is the actual wind speed, y_{max} is the maximum actual wind speed, and n is the sample size.

Tables 7–12 summarize the results of different models based on these two evaluation metrics. Tables 7 and 8, respectively, list the *NMAE* and *NRMSE* of component single-algorithm models that are trained without the feature selection procedure. Tables 9 and 10 show the evaluation results of the developed two-layer multi-model framework without the feature selection procedure. Tables 11 and 12 illustrate the results of the developed framework with the feature selection procedure.

As shown in Tables 7 and 8, none of the single-algorithm models performs better than the persistence method. Without considering the persistence model, no single-algorithm model is always most accurate at all seven locations. For example, GBM_g is the most accurate model at BND, and GBM_l performs the best at TBL. In addition, models with non-linear kernels are generally more accurate than those with linear-kernels. For instance, the SVR_poly model always outperforms the SVR_li models.

Comparing Tables 9 and 10 with Tables 7 and 8, the proposed multi-model forecasting framework with different blending algorithms outperforms the single-algorithm models. Even without the feature selection procedure, the two-layer models have improved the accuracy of the component models by up to 23.8% based on *NMAE* and 25.6% based on *NRMSE*. For the blending algorithms, the models with non-linear blending algorithms have better performance than the models with linear blending algorithms. This shows that the forecasts produced from the first-layer models exhibit a non-linear relationship with the actual wind speed. The model with the polynomial-kernel SVM algorithm is the most accurate model among all the ensemble models.

The results of the developed hybrid models in conjunction with the deep feature selection procedure are listed in Tables 11 and 12. Comparing the same hybrid models as shown in Tables 9 and 10, the deep feature selection has improved the forecasting accuracy by up to 21.86% and 19.92% based on *NMAE* and *NRMSE*, respectively. Similar to the models without feature selection, the SVM with polynomial kernel is the best performing algorithm among all hybrid algorithms. Comparing the developed multi-model framework with deep feature selection with the single-algorithm models, the multi-model forecasting model has improved the accuracy by up to 30.87% and 30.03% based on *NMAE* and *NRMSE*, respectively. In addition, this best hybrid model is performing better than the persistence forecasts at all seven locations.

3.5. Probabilistic results of the multi-model forecasting

In addition to deterministic forecasts, the developed multi-model methodology can also produce probabilistic forecasts. Fig. 9 provides an example of the deterministic forecasts along with the confidence intervals in the form of fan chart, at all seven locations. The confidence bands are calculated based on the component models. The colors of the intervals fade with the increasing confidence level, ranging from 10% to 90% in a 10% increments. The intervals are symmetric around the deterministic forecasting curves with a changing width. When the wind speed fluctuates within a small range, the confidence bands are narrow, as shown by hours 0–20 at the BND site and hours 40–50 at the DRA site. When there is a significant ramp, the uncertainty of the forecasts is increased and the bands tend to be broader, as shown by hours 20–35 at the BND site. This further proves the necessity of probabilistic forecasting.

To quantify the probabilistic forecasting accuracy, two metrics are used: reliability and sharpness. Reliability is the correct degree of a probabilistic forecasting [52], which can be assessed by the hit percentage [53]. Sharpness is the uncertainty conveyed by the probabilistic forecasts, which can be computed as the average interval size of different confident levels [54]. These two metrics are visualized by the reliability diagram and δ-diagram. In our case, the reliability diagrams of all seven locations are depicted in Fig. 10. The black solid line represents the ideal probability of the forecasts. The probabilistic forecasting produced by the developed multi-model framework is under-confident at BND and SXF, and over confident at TBL, GCM, PSU, and DRA. The reliability of the probabilistic forecasts at FPK is the best due to the smallest devia-

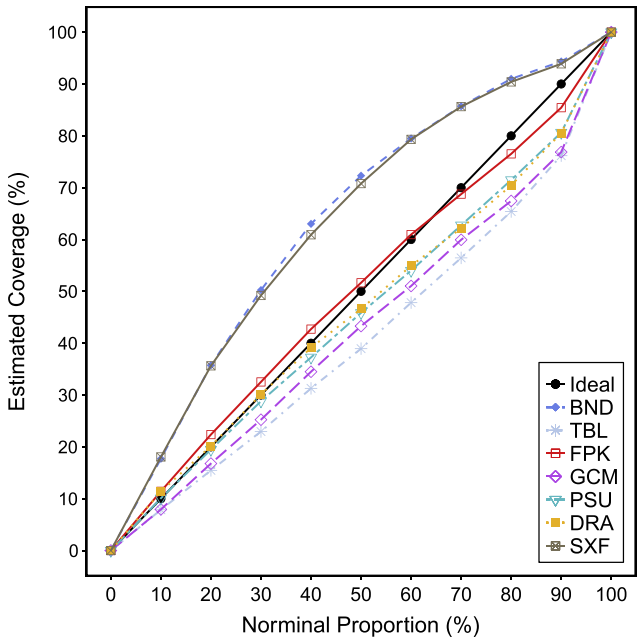


Fig. 10. Reliability diagram for the probabilistic forecasting results at seven locations.

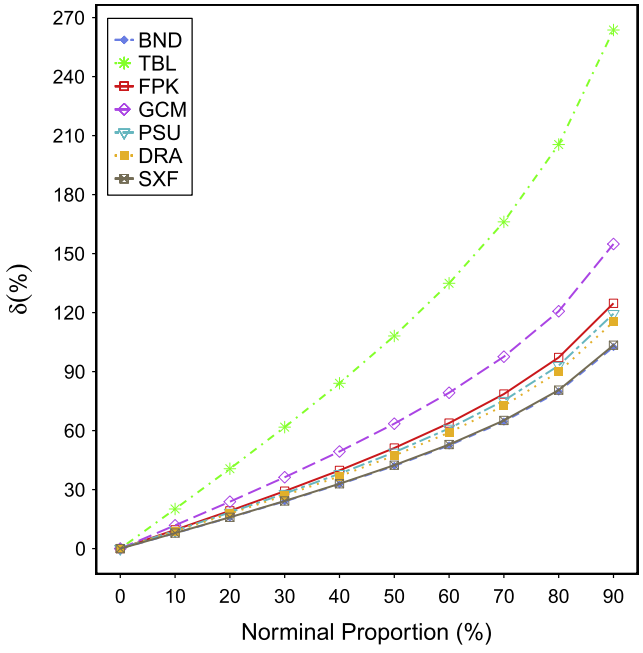


Fig. 11. δ-diagram for the probabilistic forecasting results at seven locations.

tion to the ideal line. Similarly, Fig. 11 depicts the δ-diagrams of the probabilistic forecasting at all seven location. Except for TBL, the model shows a consistent sharpness over different locations. The high sharpness of the probabilistic forecasts at TBL is mainly due to the frequent ramps of wind speed.

4. Conclusion

In this paper, a novel two-layer hybrid WPF/WSF methodology in conjunction with deep feature selection was developed. The framework consists of multiple single machine learning algorithms in the first layer and blending algorithms in the second layer. Several algorithms such as ANN, SVM, GBM, and RF with different kernels were tested and tuned in both layers. A deep feature selection framework was also developed to optimally determine the most suitable input combinations for the forecasting models. In the deep feature selection framework, the PCA, GCT, ACF, PACF, and RFE methods were adopted and implemented in an optimized sequence to take advantage of each method. The multi-model wind forecasting methodology was evaluated using data from seven SURFRAD locations. Both the hybrid algorithms and the feature selection approach were found to significantly improve the 1-h-ahead forecasting performance. The developed multi-model methodology outperformed the benchmark models by up to 30.87% and 30.03% at the 1-h-ahead forecasting horizon based on *NMAE* and *NRMSE*, respectively. Also, probabilistic forecasting produced by the developed method quantified the uncertainty of the forecasts along with the deterministic forecasting.

The developed enhanced deterministic and probabilistic wind forecasting could benefit power system operators, energy traders, and wind plant owners from different perspectives. The system operators can apply the improved forecasts in the real time security constrained unit commitment and real time security constrained economic dispatch to (i) start up/shut down generators in response to fluctuations; (ii) reduce the utilization of fast acting but expensive units; (iii) decrease the reserve levels; and (iv) reduce the wind curtailment. Overall, the improved wind forecasts would be helpful in reducing the operation costs and increasing the system reliability. The forecasts can also be used to determine the charge and discharge schedule of energy storage in a microgrid system with distributed wind generators and energy storage.

The potential future work is to validate the effectiveness of the developed multi-model framework with different time horizons in short-term forecasting. Additionally, other variables, such as temperature and humidity, can also be forecasted by first-layer models and blended by the second-layer algorithms.

Acknowledgement

This work was supported by the National Renewable Energy Laboratory under Subcontract No. XGJ-6-62183-01 (under the U. S. Department of Energy Prime Contract No. DE-AC36-08GO28308).

References

[1] Sahu BK, Hiloidhari M, Baruah D. Global trend in wind power with special focus on the top five wind power producing countries. Renew Sustain Energy Rev 2013;19:348–59. http://dx.doi.org/10.1016/j.rser.2012.11.027.
[2] Wang J, Botterud A, Bessa R, Keko H, Carvalho L, Issicaba D, Sumaili J, Miranda V. Wind power forecasting uncertainty and unit commitment. Appl Energy 2011;88(11):4014–23.
[3] Cui M, Zhang J, Florita AR, Hodge B-M, Ke D, Sun Y. An optimized swinging door algorithm for identifying wind ramping events. IEEE Trans Sustain Energy 2016;7(1):150–62. http://dx.doi.org/10.1109/tste.2015.2477244.
[4] Mendes J, Sumaili J, Bessa R, Keko H, Miranda V, Botterud A, et al. Very short-term wind power forecasting: state-of-the-art. Tech. rep.. Argonne National Laboratory (ANL); 2014.
[5] Al-Yahyai S, Charabi Y, Gastli A. Review of the use of numerical weather prediction (NWP) models for wind energy assessment. Renew Sustain Energy Rev 2010;14(9):3192–8. http://dx.doi.org/10.1016/j.rser.2010.07.001.
[6] Wedi NP, Smolarkiewicz PK. A framework for testing global non-hydrostatic models. Quart J Roy Meteorol Soc 2009;135(639):469–84. http://dx.doi.org/10.1002/qj.377.
[7] Baidya Roy S, Traiteur J, Callicutt D, Smith M. A short-term ensemble wind speed forecasting system for wind power applications. AGU fall meeting abstracts, vol. 1. p. 0850.
[8] Côté J, Gravel S, Méthot A, Patoine A, Roch M, Staniforth A. The operational cmc-mrb global environmental multiscale (GEM) model. Part I: Design considerations and formulation. Monthly Weather Rev 1998;126(6):1373–95.
[9] Jensen DG, Petersen C, Rasmussen MR. Assimilation of radar-based nowcast into a HIRLAM NWP model. Meteorol Appl 2015;22(3):485–94. http://dx.doi.org/10.1002/met.1479.

[10] Fischer C, Montmerle T, Berre L, Auger L, Ştefănescu SE. An overview of the variational assimilation in the ALADIN/France numerical weather-prediction system. Quart J Roy Meteorol Soc 2005;131(613):3477–92. http://dx.doi.org/10.1256/qj.05.115.

[11] Pichelli E, Ferretti R, Cimini D, Panegrossi G, Perissin D, Pierdicca N, Rocca F, Rommen B. InSAR water vapor data assimilation into mesoscale model MM5: technique and pilot study. IEEE J Select Top Appl Earth Observ Remote Sensing 2015;8(8):3859–75. http://dx.doi.org/10.1109/jstars.2014.2357685.

[12] Wagenbrenner NS, Forthofer JM, Lamb BK, Shannon KS, Butler BW. Downscaling surface wind predictions from numerical weather prediction models in complex terrain with WindNinja. Atmos Chem Phys Discuss 2016;16(8):5229–41. http://dx.doi.org/10.5194/acp-2015-761.

[13] Hu Q, Su P, Yu D, Liu J. Pattern-based wind speed prediction based on generalized principal component analysis. IEEE Trans Sustain Energy 2014;5(3):866–74. http://dx.doi.org/10.1109/tste.2013.2295402.

[14] Erdem E, Shi J. ARMA based approaches for forecasting the tuple of wind speed and direction. Appl Energy 2011;88(4):1405–14. http://dx.doi.org/10.1016/j.apenergy.2010.10.031.

[15] Liu H, Erdem E, Shi J. Comprehensive evaluation of arma–garch (-m) approaches for modeling the mean and volatility of wind speed. Appl Energy 2011;88(3):724–32.

[16] Silaghi H, Costea C. Wind speed prediction using Box-Jenkins method. J Comput Sci Control Syst 2008(1):208.

[17] Poncela M, Poncela P, Perán JR. Automatic tuning of kalman filters by maximum likelihood methods for wind energy forecasting. Appl Energy 2013;108:349–62. http://dx.doi.org/10.1016/j.apenergy.2013.03.041.

[18] Carpinon A, Langella R, Testa A, Giorgio M. Very short-term probabilistic wind power forecasting based on markov chain models. In: 2010 IEEE 11th international conference on probabilistic methods applied to power systems. p. 107–12. http://dx.doi.org/10.1109/pmaps.2010.5528983.

[19] Song Z, Jiang Y, Zhang Z. Short-term wind speed forecasting with markov-switching model. Appl Energy 2014;130:103–12.

[20] Ghorbani M, Khatibi R, FazeliFard M, Naghipour L, Makarynskyy O. Short-term wind speed predictions with machine learning techniques. Meteorol Atmos Phys 2016;128(1):57–72.

[21] Chitsaz H, Amjady N, Zareipour H. Wind power forecast using wavelet neural network trained by improved clonal selection algorithm. Energy Convers Manage 2015;89:588–98. http://dx.doi.org/10.1016/j.enconman.2014.10.001.

[22] Li G, Shi J. On comparing three artificial neural networks for wind speed forecasting. Appl Energy 2010;87(7):2313–20.

[23] Zhou J, Shi J, Li G. Fine tuning support vector machines for short-term wind speed forecasting. Energy Convers Manage 2011;52(4):1990–8. http://dx.doi.org/10.1016/j.enconman.2010.11.007.

[24] Troncoso A, Salcedo-Sanz S, Casanova-Mateo C, Riquelme J, Prieto L. Local models-based regression trees for very short-term wind speed prediction. Renew Energy 2015;81:589–98. http://dx.doi.org/10.1016/j.renene.2015.03.071.

[25] Liu H, Tian H-q, Pan D-f, Li Y-f. Forecasting models for wind speed using wavelet, wavelet packet, time series and artificial neural networks. Appl Energy 2013;107:191–208.

[26] Liu H, Tian H-q, Liang X-f, Li Y-f. Wind speed forecasting approach using secondary decomposition algorithm and elman neural networks. Appl Energy 2015;157:183–94.

[27] Chen K, Yu J. Short-term wind speed prediction using an unscented kalman filter based state-space support vector regression approach. Appl Energy 2014;113:690–705.

[28] Wang J-Z, Wang Y, Jiang P. The study and application of a novel hybrid forecasting model–a case study of wind speed forecasting in china. Appl Energy 2015;143:472–88.

[29] Freedman JM, Manobianco J, Schroeder J, Ancell B, Brewster K, Basu S, et al. The wind forecast improvement project (WFIP): a public/private partnership for improving short term wind energy forecasts and quantifying the benefits of utility operations. the southern study area, final report. Tech. rep.; Apr 2014. doi:http://dx.doi.org/10.2172/1129905.

[30] Xiao L, Wang J, Dong Y, Wu J. Combined forecasting models for wind energy forecasting: a case study in china. Renew Sustain Energy Rev 2015;44:271–88. http://dx.doi.org/10.1016/j.rser.2014.12.012.

[31] Liu H, Tian H-Q, Chen C, fei Li Y. A hybrid statistical method to predict wind speed and wind power. Renew Energy 2010;35(8):1857–61. http://dx.doi.org/10.1016/j.renene.2009.12.011.

[32] Liu H, qi Tian H, fei Li Y. Four wind speed multi-step forecasting models using extreme learning machines and signal decomposing algorithms. Energy Convers Manage 2015;100:16–22. http://dx.doi.org/10.1016/j.enconman.2015.04.057.

[33] Kaur A, Pedro HT, Coimbra CF. Ensemble re-forecasting methods for enhanced power load prediction. Energy Convers Manage 2014;80:582–90. http://dx.doi.org/10.1016/j.enconman.2014.02.004.

[34] Liu D, Niu D, Wang H, Fan L. Short-term wind speed forecasting using wavelet transform and support vector machines optimized by genetic algorithm. Renew Energy 2014;62:592–7. http://dx.doi.org/10.1016/j.renene.2013.08.011.

[35] Kou P, Liang D, Gao F, Gao L. Probabilistic wind power forecasting with online model selection and warped gaussian process. Energy Convers Manage 2014;84:649–63. http://dx.doi.org/10.1016/j.enconman.2014.04.051.

[36] Li S, Wang P, Goel L. Wind power forecasting using neural network ensembles with feature selection. IEEE Trans Sustain Energy 2015;6(4):1447–56. http://dx.doi.org/10.1109/tste.2015.2441747.

[37] Kong X, Liu X, Shi R, Lee KY. Wind speed prediction using reduced support vector machines with feature selection. Neurocomputing 2015;169:449–56. http://dx.doi.org/10.1016/j.neucom.2014.09.090.

[38] Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 1969;37(3):424. http://dx.doi.org/10.2307/1912791.

[39] Sun W, Liu M, Liang Y. Wind speed forecasting based on feemd and lssvm optimized by the bat algorithm. Energies 2015;8(7):6585–607.

[40] Wu Q, Peng C. Wind power grid connected capacity prediction using lssvm optimized by the bat algorithm. Energies 2015;8(12):14346–60.

[41] Granger CW. Time series analysis, cointegration, and applications. Am Econ Rev 2004;94(3):421–5. http://dx.doi.org/10.1257/0002828041464669.

[42] Fuller WA. Introduction to statistical time series, vol. 428. John Wiley & Sons; 2009.

[43] Lu S, Hwang Y, Khabibrakhmanov I, Marianno FJ, Shao X, Zhang J, Hodge B-M, Hamann HF, et al. Machine learning based multi-physical-model blending for enhancing renewable energy forecast - improvement via situation dependent error correction. In: 2015 European control conference (ECC). http://dx.doi.org/10.1109/ecc.2015.7330558.

[44] Wilczak JM, Benjamin S, Calvert S, Stern A, DiMego G, White A. Enhancing short term wind energy forecasting for improved utility operations: technical description of a joint doe/noaa/private industry collaborative field program, 2011.

[45] Bouzgou H, Benoudjit N. Multiple architecture system for wind speed prediction. Appl Energy 2011;88(7):2463–71.

[46] Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. Chemometr Intell Lab Syst 2006;83(2):83–90.

[47] Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20(3):273–97. http://dx.doi.org/10.1007/bf00994018.

[48] Silva L. A feature engineering approach to wind power forecasting. Int J Forecast 2014;30(2):395–401. http://dx.doi.org/10.1016/j.ijforecast.2013.07.007.

[49] Ibarra-Berastegi G, Saénz J, Esnaola G, Ezcurra A, Ulazia A, et al. Short-term forecasting of the wave energy flux: analogues, random forests, and physics-based models. Ocean Eng 2015;104:530–9. http://dx.doi.org/10.1016/j.oceaneng.2015.05.038.

[50] Liu D, Wang J, Wang H. Short-term wind speed forecasting based on spectral clustering and optimised echo state networks. Renew Energy 2015;78:599–608.

[51] Huang C-Y, Chiang B-Y, Chang S-Y, Tzeng G-H, Tseng C-C. Predicting of the short term wind speed by using a real valued genetic algorithm based least squared support vector machine. In: Intelligent decision technologies. Springer; 2011. p. 567–75.

[52] Taylor JW, Jeon J. Forecasting wind power quantiles using conditional kernel estimation. Renew Energy 2015;80:370–9.

[53] Pinson P, Møller JK, Nielsen HA, Madsen H, Kariniotakis GN. Evaluation of nonparametric probabilistic forecasts of wind power. Tech. rep.. DTU: Informatics and Mathematical Modelling, Technical University of Denmark; 2007.

[54] Gallego-Castillo C, Bessa R, Cavalcante L, Lopez-Garcia O. On-line quantile regression in the rkhs (reproducing kernel hilbert space) for operational probabilistic forecasting of wind power. Energy 2016;113:355–65.