Wind Power Forecasting Using Neural Network Ensembles With Feature Selection

Li, Song; Wang, Peng; Goel, Lalit

2015

Li, S., Wang, P., & Goel, L. (2015). Wind Power Forecasting Using Neural Network Ensembles With Feature Selection. IEEE Transactions on Sustainable Energy, 6(4), 1447-1456.

https://hdl.handle.net/10356/80608

https://doi.org/10.1109/TSTE.2015.2441747

Wind Power Forecasting Using Neural Network Ensembles with Feature Selection

Song Li, Peng Wang, *Senior Member, IEEE*, and Lalit Goel, *Fellow, IEEE*

***Abstract*—In this paper, a novel ensemble method consisting of neural networks, wavelet transform, feature selection and partial least squares regression is proposed for the generation forecasting of a wind farm. Based on the conditional mutual information, a feature selection technique is developed to choose a compact set of input features for the forecasting model. In order to overcome the nonstationarity of wind power series and improve the forecasting accuracy, a new wavelet-based ensemble scheme is integrated into the model. The individual forecasters are featured with different mixtures of the mother wavelet and the number of decomposition levels. The individual outputs are combined to form the ensemble forecast output using the partial least squares regression method. To confirm the effectiveness, the proposed method is examined on real-world datasets and compared to other forecasting methods.**

***Index Terms*—Feature selection, neural networks, partial least squares regression, wavelet transform, wind power forecasting.**

I. Introduction

WIND power as a sustainable energy source is becoming a promising supplement for electric power generation. By the end of 2013, the installed commercial wind power in more than 90 countries reached a total capacity of 318 GW, offering about 3% of global electricity [1]. However, the uncertainties of wind power make power system operation and control more challenging. An accurate and reliable wind power forecasting (WPF) model is therefore essential for spinning reserve, power quality and reliability management [2].

Many WPF techniques have been proposed in the literature, which can be roughly classified into three categories: physical approach, statistical approach and a combination of both. The physical method uses topological and meteorological variables as the inputs to produce the best possible wind speed forecasts. The predicted wind speeds are then converted into wind power forecasts via the power curves of wind turbines. The statistical method tries to develop a relationship between the future wind power and a set of variables such as historical wind power data and historical and forecasted wind speed and direction values. The time series models like autoregressive moving average (ARMA) and Kalman filtering [3]-[5] and the artificial intelligence models including neural networks, fuzzy logic and support vector machines [6]-[10] are examples of the statistical method. The physical method has merits in long-term forecasts (6 hours ahead and more) while the statistical method performs well in short-term [11]. In addition, the hybrid of physical and statistical approaches is also an alternative for WPF, which can fully use their respective advantages [12], [13].

In particular, neural networks (NNs) based forecasters have obtained many good results in WPF [14], but there is still a lot of potential for improvements. First, the performance of NN is sensitive to the initialization of weights and biases. Second, the overtraining problem may arise if NN has too many parameters to be estimated from the training data. An effective method for overcoming the above problems is to propose a neural network ensemble for WPF.

A neural network ensemble consists of multiple NNs, which are all attempting to solve the same WPF problem. It is shown that an ensemble of forecasters has advantages over individual forecasters in terms of increased accuracy and robustness. The improvements can be attributed to that the diverse errors of the individual forecasters randomly scatter on the different parts of input space and can cancel out in the aggregating process [15]. Many ensemble methods have been reported for WPF [16]-[19]. In [16], weather ensemble predictions were converted to wind power density forecasts through the power curve. An ensemble model of 52 NN sub-models and 5 Gaussian Process (GP) sub-models was presented in [17], in which the NNs were used to train the wind information and the GPs were used to provide the initial power level to the NN sub-models.

The wavelet transform has been successfully used for WPF by breaking up the wind power series into a set of constitutive components [7], [20]. However, it actually can play a bigger role in the modeling process. In this paper, a novel wavelet-based ensemble scheme is proposed to create a neural network ensemble model for WPF. It is known that the two key wavelet parameters are needed during the transform: the type of mother wavelet and the number of decomposed levels. In the proposed ensemble strategy, each individual forecaster corresponds to a different combination of the mother wavelet and the number of decomposition levels. The motivation is that different wavelet parameters will result in different input data (i.e. varying initial conditions) for the NN-based forecasters, which can encourage the error diversity of the neural network ensemble.

In ensemble forecasting, the ensemble forecast should be a weighted combination of individual forecasts, because some of the individuals are actually more precise than others. Since the individual predictors are used to solve the same WPF problem, their forecast outputs will be highly correlated. In this case, the multiple linear regression (MLR) method may come across the

Song Li, Peng Wang and Lalit Goel are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. (e-mail: sli5@e.ntu.edu.sg; epwang@ntu.edu.sg; elkgoel@ntu.edu.sg).

collinearity or singularity problem and yield inaccurate weight factors. In this paper, partial least squares regression (PLSR) is employed to compute the weight factors, which can remedy the drawbacks in MLR [21].

In the development of a WPF model, selecting a suitable set of input variables from the raw data has a great impact on the forecasting performance. Generally, more input variables carry more discriminating power, but in practice, excessive variables are prone to cause many problems, such as prolonged training period and the curse of dimensionality [22]. The reason is that some variables are irrelevant or redundant to the model, which would confuse the model and result in poor prediction results. This paper describes a conditional mutual information based feature selection (CMIFS) approach to determine a small set of informative input variables that are beneficial to WPF.

In this paper, a novel ensemble method consisting of neural networks, wavelet transform, feature selection and partial least squares regression is proposed to forecast the wind generation. The main contributions of this paper are as follows:

1) A new ensemble model for WPF is proposed, which can alleviate the uncertainty and overtraining problems and improve the forecasting accuracy and robustness.
2) CMIFS is developed to assist the WPF model, which can carry out relevance and redundancy analysis and select a compact set of informative input variables efficiently.
3) Wavelet transform is not only utilized to decompose the wind power data but also to generate the neural network ensemble. The proposed ensemble scheme can avoid the selection process of perfect wavelet parameters, suppress the biases induced by fix parameters and make use of the independent information of imperfect parameters.
4) PLSR is introduced as a combination method to form the ensemble forecast, which can identify the contribution of each individual forecast and handle the collinearity issue. Furthermore, PLSR allows the selection of the number of latent components used in the regression process, which can also reduce the chance of overtraining.

The proposed WPF method is tested on real-world data and compared to other forecasting methods. Section II presents the relevant theories and describes the proposed method in details. Section III provides the numerical results and discussions. The conclusion of this paper is outlined in Section IV.

II. Proposed Wind Power Forecasting Method

A. Conditional Mutual Information Based Feature Selection

Let X be a random variable with the probability distribution function $p(x)$, the entropy $H(X)$ is defined by

$$H(X) = -\sum_{x} p(x)\log p(x). \tag{1}$$

The conditional entropy $H(X|Y)$ is defined by

$$H(X \mid Y) = -\sum_{x}\sum_{y} p(x,y)\log p(x \mid y) \tag{2}$$

where $p(x,y)$ is the joint probability distribution of X and Y and $p(x|y)$ is the conditional distribution of X given the outcome of Y. The entropy $H(X)$ is a measure of the amount of uncertainty about X, while the conditional entropy $H(X|Y)$ states how much uncertainty remains in X after Y is known [23], [24].

The mutual information (MI) between X and Y is defined by

$$I(X;Y) = \sum_{x}\sum_{y} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)}. \tag{3}$$

The MI gives an estimate of the quantity of dependence of two variables. Large MI implies that the two variables are strongly correlated and vice versa. The MI can be rewritten by

$$I(X;Y) = H(X) - H(X \mid Y) \tag{4}$$

which indicates that $I(X;Y)$ quantifies how much uncertainty of X is reduced due to the outcome of Y. The conditional mutual information (CMI) between X and Y given Z is defined by

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z). \tag{5}$$

$I(X;Y|Z)$ measures the amount of information shared between X and Y when Z is known. If Y and Z bring the same information about X, $H(X|Z)$ and $H(X|Y,Z)$ are equal and $I(X;Y|Z)$ is zero. On the contrary, if Y contains information about X that Z does not have, the difference on the right is large and so is $I(X;Y|Z)$.

Let C be the dependent variable, $F = \{F_1, F_2, \ldots, F_N\}$ be the candidate set with N features, the goal of feature selection is to choose a subset $S \subset F$ with K features such that the conditional entropy $H(C|S)$ is minimized. But in practice, the minimization of $H(C|S)$ is computationally intractable as the total number of subsets for evaluation is prohibitive [25]. Hence, approximated solutions are developed for feature selection in applications.

Fleuret proposed a fast feature selection technique based on the CMI, which can ensure that the selected features are both individually informative and mutually weakly dependent [26]. In this technique, the first selected feature is the one that holds the maximum MI with C. The new feature F_j is regarded as a good one if it carries information about C and this information has not been caught by the selected features. At this stage, the selection criterion relies on how much information F_j can add with respect to the existing features. The next feature F_j to be selected is the one that makes $I(C;F_j|F_i)$ large for every feature F_i in the selected subset S. Following this scheme, the feature is chosen one by one until S is full of K features.

The above feature selection approach has two problems that need to be addressed. First, it is possible to select bad features if only employing the measure $I(C;F_j|F_i)$. Based on the concept of trivariate mutual information, $I(C;F_j|F_i)$ can be expressed by

$$I(C;F_j \mid F_i) = I(C;F_j) - I(C;F_j;F_i) \tag{6}$$

where $I(C;F_j;F_i)$ measures the amount of information shared by C, F_j and F_i. It is found that the large value $I(C;F_j|F_i)$ could be obtained by two small values $I(C;F_j)$ and $I(C;F_j;F_i)$. The small $I(C;F_j)$ indicates that the new feature F_j is weakly correlated to C. To overcome this problem, a preliminary selection process for relevant features is included. In this step, the MI values are calculated for every candidate feature and ranked from high to low. Then the features with top T positions are retained and the

others are removed from the candidate set. The new set is used for further selection by the Fleuret's approach. Note that T is a threshold parameter for filtering out the irrelevant and weakly relevant features. In the literature, the number of inputs used for WPF is at most twenty or thirty. Thus, the threshold T is set to be forty in this paper, which is deemed enough to include all relevant features.

Secondly, it is clear that the estimate of MI values is crucial for the efficacy of feature selection. In [26], Fleuret's approach was actually proposed for binary feature selection in a context of classification. However, the inputs and outputs in WPF take continuous values, making the MI more difficult to estimate. In this paper, the estimation method proposed in [23] is adopted. This estimator utilizes the merit of Parzen windows, which can calculate the MI accurately and efficiently.

In this paper, for simplicity, only the wind power and wind speed measurements are considered as the candidate inputs for the forecaster. The wind power is the total generation output of a wind farm and the wind speed is measured at the site of wind farm. Hence, a candidate set of input features, denoted by $F(t)$, is presented by

$$F(t) = \{\mathrm{WP}(t-1), ..., \mathrm{WP}(t-100), \mathrm{WS}(t), ..., \mathrm{WS}(t-100)\} \quad (7)$$

where WP(t-1),…,WP(t-100) and WS(t),…,WS(t-100) are the wind power and speed data up to 100 hours ago, respectively. Since the wind power data does not exhibit apparent recurring trends, the order of historical values is set to be 100, which is sufficient to contain all useful features. Similar thing happens to the wind speed series, too. The candidate set $F(t)$ totally has 201 features, which cannot be directly used in WPF.

Assuming that the candidate set F consists of N features and the subset S to be determined contains K features, the proposed feature selection procedure can be summarized as follows:

1) Obtain the target variable C (i.e. wind power vector) and its corresponding candidate set F of N (=201) features.
2) For each feature $F_j \in F$, compute $I(C;F_j)$ and rank them.
3) Move the top T (=40) features to a new set F_{new}.
4) Initialize the selected subset $S = \{\emptyset\}$.
5) Locate the feature F_k in F_{new} that maximizes $I(C;F_k)$. Set $F_{new} \leftarrow F_{new} \setminus \{F_k\}$ and $S \leftarrow S \cup \{F_k\}$.
6) For all pairs of features (F_i, F_j) with $F_i \in S$ and $F_j \in F_{new}$, compute $I(C;F_j|F_i)$.
7) Select the next feature F_k, which is defined by

$$F_k = \arg \max_{F_j \in F_{new}} \min_{F_i \in S} \left\{ I(C; F_j \mid F_i) \right\}. \quad (8)$$

8) Set $F_{new} \leftarrow F_{new} \setminus \{F_k\}$ and $S \leftarrow S \cup \{F_k\}$.
9) Repeat steps 6) to 8) until S is full of K features.
10) Output the subset S containing the selected features.

B. Wavelet Neural Networks

In this paper, three-layer feedforward neural networks with Levenberg-Marquardt (LM) learning algorithm are adopted as the forecast engine. LM inherits the speed advantage of Gauss-Newton method and the stability advantage of steepest descent method, leading to a very efficient and robust learning process [27]. To improve the generalization capability and refrain from the overtraining problem, the early stopping method [27] and the Nguyen-Widrow initialization method [28] are used in the forecasting model.

In pursuit of better accuracy, it is necessary to probe deeply into the internal characteristics of wind power series. It is seen that the wind power series has multiple frequency components, which are always the challenging parts in WPF. In this paper, wavelet transform is used to cut up the wind power data into a set of constitutive components that have different frequencies. Each constitutive component is forecasted by a NN. Compared to the original wind power series, the constitutive components present better behaviors (e.g. more stable variance and fewer outliers) and therefore can be forecasted more accurately [9].

Wavelet analysis makes use of two basic functions: mother wavelet function and scaling function. Given a mother wavelet $\psi(t)$ and its corresponding scaling function $\varphi(t)$, a wind power series $w(t)$ can be represented by

$$\begin{aligned} w(t) = &\sum_k c_{j_0}(k) 2^{j_0/2} \varphi\left(2^{j_0} t - k\right) \\ &+ \sum_k \sum_{j=j_0}^{\infty} d_j(k) 2^{j/2} \psi\left(2^j t - k\right) \end{aligned} \quad (9)$$

where t is the time index, j_0 is the predefined scale, j and k are integer variables for scaling and translation and $c_{j0}(k)$ and $d_j(k)$ are the approximation and detail coefficients, respectively. The coefficients $c_{j0}(k)$ and $d_j(k)$ in this representation are called the discrete wavelet transform of $w(t)$ [29].

The multilevel wavelet decomposition is a process to obtain "approximations" and "details" based on the given signal. An approximation is a coarse representation of the original signal, whereas a detail component is the difference of two successive approximations [30]. An example of two-level decomposition for the wind power series is given by

$$w(t) = A_1(t) + D_1(t) = A_2(t) + D_2(t) + D_1(t). \quad (10)$$

The wind power series w is firstly cut up into two components: A_1 and D_1. Then the approximation A_1 is further broken up into two components: A_2 and D_2. The approximation component A_2 reflects the general trend and presents a smooth form of w. The terms D_2 and D_1 describe the high frequency components in w.

C. Wavelet-Based Ensemble Scheme

Although many papers have applied the wavelet transform to WPF, there is no fixed criterion concerning the selection of two parameters: mother wavelet and number of decomposition levels. Some papers have tested many wavelet parameters and selected the most suitable combination based on the prediction results. However, this selection process presents at least three shortcomings. First, the testing process of numerous wavelet specifications would take too much time. Only the best setting is selected and the others are thrown away directly, which is actually a kind of waste. Second, the fixed specification may not always represent the wind power series adequately because the wind power series changes all the time. Third, the optimal

wavelet specification for one look-ahead time may not be still good for other look-ahead times, which is shown in Case 4.

In this paper, a wavelet-based ensemble scheme is proposed to overcome the above problems. The individual wavelet NNs are featured with different combinations of the mother wavelet and the number of decomposition levels. To select the suitable wavelet specifications, the attributes of mother wavelet and the characteristics of wind power data must be taken into account [30]. Daubechies (db), Coiflets (coif) and Symlets (sym) are compactly supported wavelets with high number of vanishing moments [29]. These properties can result in an appropriate trade-off between wavelength and smoothness, which are very suitable for treating the nonstationary wind power series [9]. In fact, the family of Daubechies wavelets has been widely used to cut up the wind power data [9], [20]. In this paper, we expand to two more wavelet families: Coiflets and Symlets. The order of mother wavelet considered is from 2 to 5 and the number of decomposition levels is 1 and 2. Hence, 12 mother wavelets (db2–db5, coif2–coif5 and sym2–sym5) are employed and the proposed wavelet-based ensemble strategy contains 24 sets of wavelet parameters. It is noted that more individual forecasters can be included by increasing the number of decomposition levels. Following this scheme, the diversity of ensemble model will be amplified and more accurate forecasting performance can be expected. However, the computing time would increase dramatically.

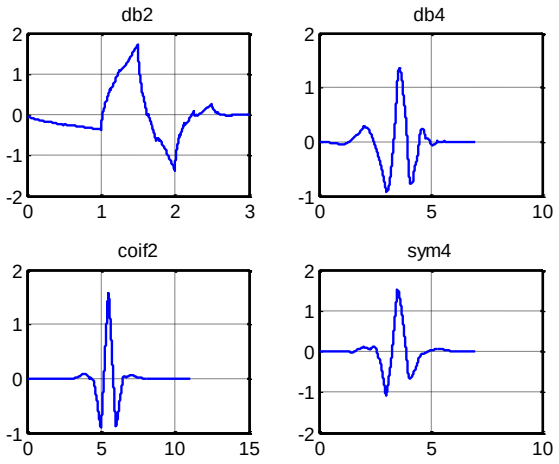


Fig. 1. Four mother wavelet functions.

Fig. 1 shows four mother wavelet functions: db2, db4, coif2 and sym4. It is seen that a wavelet is a wave-like oscillation in a short time. As the wavelets are in different trajectories, their corresponding constitutive components will also have different behaviors. For example, the four wavelets are used to carry out two-level wavelet decomposition on a wind power series. The resultant detail components D_1 are shown in Fig. 2.

Owing to the proposed wavelet-based ensemble scheme, the individual forecast is fostered based on the input variables that are different from those for the other forecast. Every individual forecast in the ensemble may contain some useful independent information and the ensemble model can take advantage of the complementary information to boost the forecasting capability. Furthermore, the ensemble model will refrain from looking for the appropriate wavelet setting and suppress the biases induced by fixed wavelet parameters.

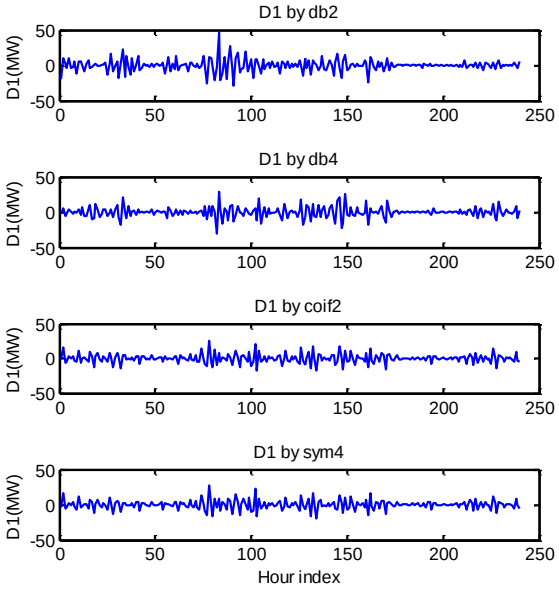


Fig.2. Detail components obtained with db2, db4, coif2 and sym4.

D. Partial Least Squares Regression

In ensemble forecasting, simple averaging is usually used to combine the individual forecasts [15], which assigns identical weight coefficients to the individuals. However, it neglects the fact that some individual forecasters are actually more accurate than the others. Considering the relative accuracy between the individual forecasts, the weighted averaging method forms the ensemble forecast by

$$z_e = \sum_{i=1}^{\beta} \alpha_i z_i \tag{11}$$

where z_e is the ensemble output, z_i is the individual output, β is the number of individual outputs and α_i is the weight factor. In this paper, PLSR is adopted to determine the weight factors.

The main idea of PLSR is to find a linear regression model by projecting the original variables to a new space. Supposing $\mathbf{X}$ and $\mathbf{Y}$ are the individual outputs and ensemble output, PLSR is going to predict $\mathbf{Y}$ from $\mathbf{X}$ through the latent variables in the new space. In order to make the regression efficient, $\mathbf{X}$ and $\mathbf{Y}$ are transformed by mean-centering and variance-scaling to $\mathbf{X_c}$ and $\mathbf{Y_c}$, respectively. Then PLSR decomposes $\mathbf{X_c}$ and $\mathbf{Y_c}$ using the outer relations [21], [31], [32]:

$$\begin{aligned} \mathbf{X_c} &= \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \sum_{j=1}^{J} \boldsymbol{t}_j \boldsymbol{p}_j^{\mathrm{T}} + \mathbf{E} \\ \mathbf{Y_c} &= \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} = \sum_{j=1}^{J} \boldsymbol{u}_j \boldsymbol{q}_j^{\mathrm{T}} + \mathbf{F} \end{aligned} \tag{12}$$

where **T** and **U** are the score matrices, **P** and **Q** are the loading matrices and **E** and **F** are the residual matrices. The superscript "T" denotes transposition and J is the number of components. PLSR intends to find the latent vectors $\boldsymbol{t}_j$ and $\boldsymbol{u}_j$ to maximize the covariance between $\mathbf{X_c}$ and $\mathbf{Y_c}$ with the condition that both the residual terms **E** and **F** are reduced [21].

It is seen that the matrices $\mathbf{X_c}$ and $\mathbf{Y_c}$ have been replaced by the new ones **T** and **U** in (12), which have smaller sizes, better properties (orthogonality) and also span the spaces of $\mathbf{X_c}$ and $\mathbf{Y_c}$. To bond $\mathbf{X_c}$ and $\mathbf{Y_c}$, PLSR builds an inner relation between **T** and **U** by

$$\mathbf{U} = \mathbf{TB} + \mathbf{U_E}. \tag{13}$$

The matrix **B** denotes the regression coefficients and $\mathbf{U_E}$ is an error term similar to **E** and **F**.

If all error terms are minimized, $\mathbf{Y_c}$ can be estimated by

$$\hat{\mathbf{Y}}_\mathbf{c} = \mathbf{UQ}^{\mathrm{T}} = \mathbf{TBQ}^{\mathrm{T}}. \tag{14}$$

The solutions to the above three equations can be computed by the SIMPLS method. The background of SIMPLS is out of the scope of this paper, but can be found in [33].

It should be noted that the number of components J in (12) is a very important parameter of a PLSR model. The maximum number of components is the rank of the matrix **X**. However, it is not proper to use as many components as possible normally. The reasons are that the observed data are never noise-free and some of the smaller components only describe the noise, which may bring the problems of collinearity and overtraining. In this paper, the cross validation method is used to select the suitable number of components. The prediction residual sum of squares (PRESS) is calculated each time a new component is involved. The number that gives a minimal PRESS value is chosen.

Due to the high correlation among the individual forecasts, the multiple linear regression (MLR) may encounter the matrix inversion issue in finding weight factors. PLSR, as a remedy for the shortcomings of MLR, conducts the regression analysis on the latent variables rather than the original variables. In this way, PLSR is able to handle collinearity and generate reliable estimates for the combining weights. Moreover, PLSR allows selecting a subset of latent components with the best estimate of generalization performance, which can get rid of overfitting (including too many components) the individual forecasts [21].

E. Proposed WPF Model

The structure of the proposed ensemble method is shown in Fig. 3. Only the wind power and wind speed data are utilized as input variables to the forecast model. The other data such as wind direction and temperature are not considered here, but can be easily involved in the future work. Note that wind speed is not displayed in Fig. 3 for simplicity.

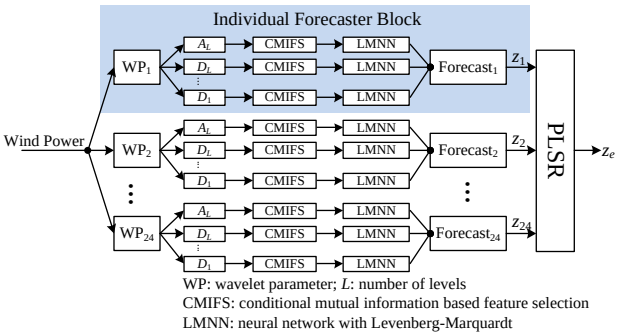


Fig. 3. Proposed ensemble model.

As shown in Fig. 3, the proposed ensemble method contains 24 individual forecaster blocks. Each forecaster block is made up of wavelet transform, input feature selection procedure and neural network (LMNN). As discussed in Section II.C, 24 sets of wavelet parameters are used for wind power decomposition. Note that wind speed data are not decomposed by the wavelet transform. In each individual forecaster, the proposed CMIFS technique is applied to the constitutive components (A_L, D_L,…, D_1). Taking A_L as an example, the candidate input set is {$A_L(t\text{-}1)$,…, $A_L(t\text{-}100)$, WS(t),…, WS(t-100)}. For each constitutive component, a three-layer LMNN is used to forecast the hourly wind power output of the next day. The Levenberg-Marquardt learning algorithm is used to train the neural networks. Finally, the 24 individual forecasts are linearly aggregated to yield the ensemble output, where the weight factors are obtained by the PLSR method. It is noted that the number of components used in PLSR is determined by the cross validation method.

III. Results and Discussions

In this section, the proposed method is tested using the data collected by the National Renewable Energy Laboratory [34]. Two different wind farms are considered for numerical testing. The first wind farm at site 1 is located in latitude 34.98° and longitude –104.04°, with installed capacity of 171.8 MW. The second wind farm at site 50 is located in latitude 47.58° and longitude –97.5°, with installed capacity of 1003.7 MW. It is clear that the two wind farms are quite different in location and installed capacity. The 10-minute wind speed and wind power measurements from 2005 to 2006 of two wind farms are used for experiment. The six readings over an hour are averaged to obtain hourly data. The site 1 wind farm data are used in Cases 1 to 5, while the site 50 wind farm data are used in Case 6.

To evaluate the performance, two different error measures: mean absolute percentage error (MAPE) and normalized root mean square error (NRMSE) [35], [36] are calculated by

$$\begin{aligned} \mathrm{MAPE} &= \frac{100\%}{M}\sum_{i=1}^{M}\left|\frac{y_i-\hat{y}_i}{y_c}\right| \\ \mathrm{NRMSE} &= \frac{100\%}{y_c}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(y_i-\hat{y}_i\right)^2} \end{aligned} \tag{15}$$

where M is the number of hours in the testing period, y_i is the actual wind power, $\hat{y}_i$ is the forecasted wind power and y_c is the

installed capacity of wind farm.

A. Forecasting Results

Case 1: It is pointed out that any advanced model for WPF should first be compared with the persistence model [37]. The persistence model states that the wind power output in the near future is the same as its last measurement, i.e.

$$\mathrm{WP}(t+\Delta t)=\mathrm{WP}(t) \tag{16}$$

where WP(t) is the wind power at time t and Δt is the forecast horizon. The persistence method performs well for very short-term forecasts but its error grows with the increasing forecast horizon. To obtain a good performance over a longer horizon, a new reference model is presented in [38], which is given by

$$\mathrm{WP}(t+\Delta t)=a\,\mathrm{WP}(t)+(1-a)\,\mathrm{WPA}(t) \tag{17}$$

where a is the correlation factor between WP(t) and WP(t+Δt) and WPA(t) is the average of the past wind power values prior to t. The two quantities WPA(t) and a are estimated by

$$\mathrm{WPA}(t)=\frac{1}{N_{\mathrm{WP}}}\sum_{t=1}^{N_{\mathrm{WP}}}\mathrm{WP}(t) \tag{18}$$

$$a=\frac{\sum_{t=1}^{N_{\mathrm{WP}}}\mathrm{WPX}(t)\mathrm{WPX}(t+\Delta t)}{\sqrt{\sum_{t=1}^{N_{\mathrm{WP}}}\mathrm{WPX}(t)^2}\sqrt{\sum_{t=1}^{N_{\mathrm{WP}}}\mathrm{WPX}(t+\Delta t)^2}} \tag{19}$$

and

$$\mathrm{WPX}(t)=\mathrm{WP}(t)-\mathrm{WPA}(t) \tag{20}$$

where N_{WP} is the predefined number of past wind power values used in calculation.

In this case, the proposed method is compared to the above two reference models using the data from the site 1 wind farm. The training set is from January 1, 2005 to December 31, 2005 and the testing set ranges from January 1, 2006 to December 31, 2006. Forty percent of data in the training set are used for validation.

The wind speeds used in the simulation are actual measured data. For each constitutive component, the number of selected input features and the number of hidden nodes are determined through extensive tests. All simulations are run in Matlab on a computer with a 2.66-GHz CPU. Note that the forecast model is not retrained in the whole testing period. But more retraining may be a good attempt to improve the forecasting accuracy.

The 1-hour to 48-hour ahead forecasting results of the three methods are shown in Table I. It is seen that the new reference model obtains better forecasting accuracy than the persistence model in 24- and 48-hour cases. The percentage increment of the proposed method with respect to the two reference models is shown in Fig. 4. On the given testing period, the proposed method demonstrates significant improvements over the other two reference methods in all look-ahead times. The proposed method is 82.5% and 77.9% better than the persistence and new reference methods in the whole range of 48 look-ahead hours in MAPE, respectively. Similar improvements have also been observed in the measure of NRMSE.

TABLE I
FORECASTING RESULTS OF PERSISTENCE, NEW REFERENCE AND PROPOSED METHOD

		1-hour	24-hour	48-hour
Persistence	MAPE	7.83	26.90	30.67
	NRMSE	12.50	35.50	39.21
New Reference	MAPE	8.30	21.48	24.37
	NRMSE	12.26	23.86	26.93
Proposed	MAPE	0.49	4.83	5.38
	NRMSE	0.66	6.29	6.98

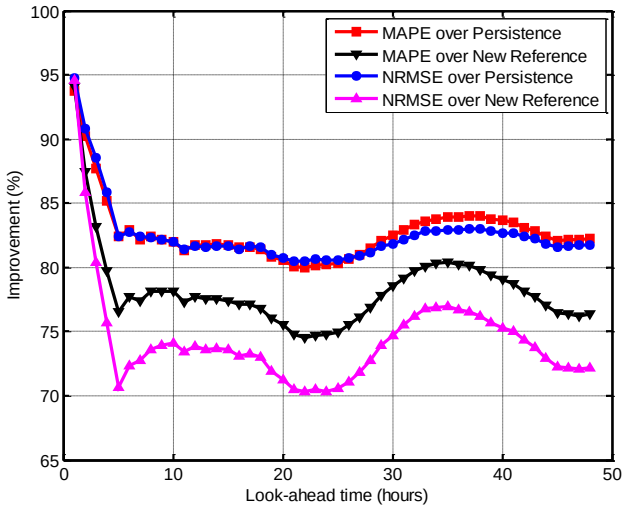


Fig. 4. Improvement of proposed method over persistence and new reference.

Case 2: This case compares the proposed method to three methods: LMNN, radial basis function NN (RBFNN) and wavelet NN (WNN) using the data in Case 1. The combination of mother wavelet coif4 and 2-level decomposition is used for WNN, which can obtain the best average forecasting accuracy for 48 look-ahead hours. The MAPE and NRMSE results of the four models are shown in Fig. 5 and Fig. 6, respectively.

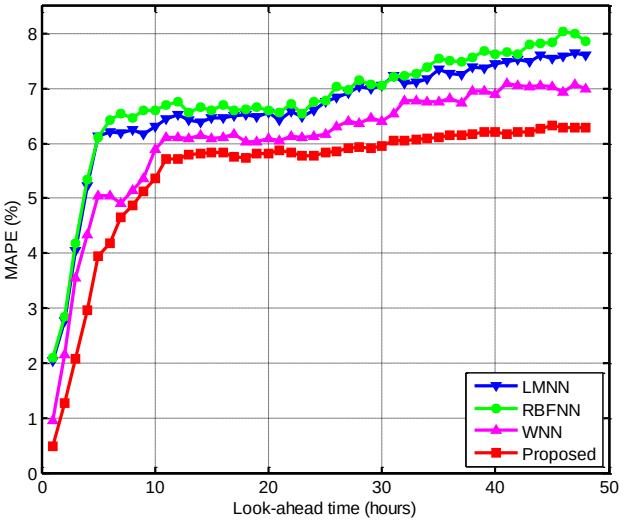


Fig. 5. MAPE results of LMNN, RBFNN, WNN and proposed method.

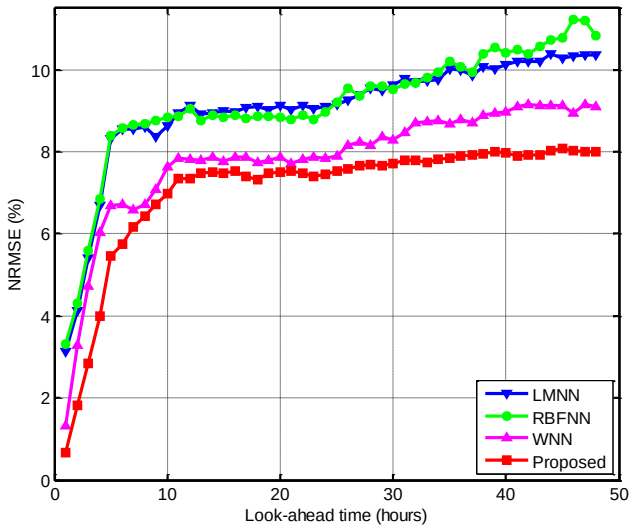


Fig. 6. NRMSE results of LMNN, RBFNN, WNN and proposed method.

It is clear that the forecasting errors of all methods increase dramatically in the first few look-ahead hours. Second, the two models LMNN and RBFNN achieve similar results, especially in short horizons. Third, the forecasting performance is clearly improved when the wavelet transform is involved. Compared to LMNN, the average accuracy of WNN is 8.48% and 13.6% better in MAPE and NRMSE, respectively. Fourth, it is shown that the proposed ensemble strategy is an effective treatment to improve the forecasting accuracy. For example, the proposed ensemble model has yielded an average reduction of 10.9% in MAPE with respect to the WNN method. Finally, the proposed method outperforms other alternatives in all look-ahead times, which confirms its effectiveness and superiority.

Case 3: In this case, various feature selection techniques are compared on the data used in Case 1. For each constitutive component (A_L, D_L,…, D_1), different feature selection methods are used to select the same number of inputs. The results for 1-hour ahead forecasting are shown in Table II. The benchmark methods in Table II include correlation analysis (CA), mutual information (MI), RReliefF [39] and minimum redundancy maximum relevance (mRMR) [40]. The first three approaches choose the features with top *scores* (e.g. correlation coefficient for CA) from the candidate input set and do not consider the redundancy problem. The mRMR method regulates the feature relevance and redundancy through comparing the new feature with the selected ones, like our proposed CMIFS method.

TABLE II
1-HOUR AHEAD FORECASTING RESULTS OF FIVE FEATURE SELECTION METHODS

	MAPE	NRMSE
CA	0.58	0.83
MI	0.62	0.91
RReliefF	0.61	0.87
mRMR	1.51	2.18
CMIFS	0.49	0.66

According to Table II, the best forecasting performance on the testing period is obtained by our proposed CMIFS method. For example, CMIFS can improve the forecasting accuracy by 15.5%, 21.0%, 19.7% and 67.5% in MAPE with respect to the preceding four methods, respectively. The improvements over CA, MI and RReliefF indicate that relevance analysis alone is insufficient for feature selection. However, the mRMR method including redundancy analysis also produces poor results. The reason could be that the redundancy test is too stringent, which removes too many strongly relevant input features.

Case 4: The forecasting potential of individual forecast and ensemble forecast is further studied in this case. The individual forecast employs a specific set of wavelet parameters while the ensemble forecast assembles the individual forecasts by PLSR. The forecasting results (in MAPE) of four different look-ahead times (1-, 6-, 12- and 24-hour ahead) are tabulated in Table III. In this table, "LAST12" refers to the ensemble forecast which only aggregates the individual forecasts associated with 2-level decomposition. "ALL24" is our proposed method using all the 24 individual forecasts.

It can be observed that different wavelet parameters lead to different forecast performance for a certain look-ahead time. A specific set of wavelet parameters cannot obtain the best result in all look-ahead times. Moreover, the two ensemble forecasts (LAST12 and ALL24) have obtained better results than any of the individual forecasts, which proves the effectiveness of the proposed wavelet-based ensemble scheme. Besides, compared to LAST12, ALL24 can produce better results even with some *bad* individual forecasts (i.e. large MAPEs). This implies that the *bad* individuals still contain some independent information that contributes to the ensemble forecast accuracy.

TABLE III
INDIVIDUAL AND ENSEMBLE FORECASTING RESULTS IN MAPE OF FOUR DIFFERENT LOOK-AHEAD TIMES

Level	Wavelet	1-hour	6-hour	12-hour	24-hour
1	coif2	1.06	5.17	5.95	6.06
	coif3	1.38	4.61	5.94	5.94
	coif4	1.18	4.79	5.98	5.98
	coif5	1.57	5.14	6.06	6.10
	sym2	2.96	4.52	6.07	6.00
	sym3	2.67	4.60	5.93	5.93
	sym4	2.09	4.68	5.97	6.04
	sym5	1.39	4.52	5.95	5.99
	db2	2.90	4.62	5.94	5.97
	db3	2.56	4.56	5.98	5.95
	db4	2.34	4.54	6.03	5.94
	db5	1.63	4.53	5.96	6.05
2	coif2	1.06	4.95	6.13	6.11
	coif3	0.86	4.94	6.03	6.28
	coif4	0.96	5.05	6.10	6.13
	coif5	0.87	5.07	6.19	6.37
	sym2	1.01	5.28	6.37	6.24
	sym3	1.74	5.23	6.10	6.33
	sym4	0.98	4.93	6.19	6.17
	sym5	0.81	4.94	6.25	6.35
	db2	1.01	5.29	6.56	6.63
	db3	1.61	4.90	6.21	6.23
	db4	1.49	4.84	6.14	6.12
	db5	0.88	5.00	6.15	6.15
LAST12		0.57	4.43	5.81	5.89
ALL24		0.49	4.19	5.71	5.78

Case 5: The preceding four case studies are made based on the measured wind speed data. In practice, the forecasted wind speed values should be included in the process of WPF, which are provided by the near weather station. In this case, the effect of wind speed forecasting errors on WPF is inspected. A noisy wind speed data series is used to simulate the prediction errors. The Gaussian noise of zero mean and standard deviation of 0.5 m/s is added to the measured wind speed series. The simulated wind speed forecasting error fluctuates between -2.16 m/s and 2.21 m/s. The 1-hour to 24-hour ahead prediction results using measured and noisy wind speed data are shown in Fig. 7.

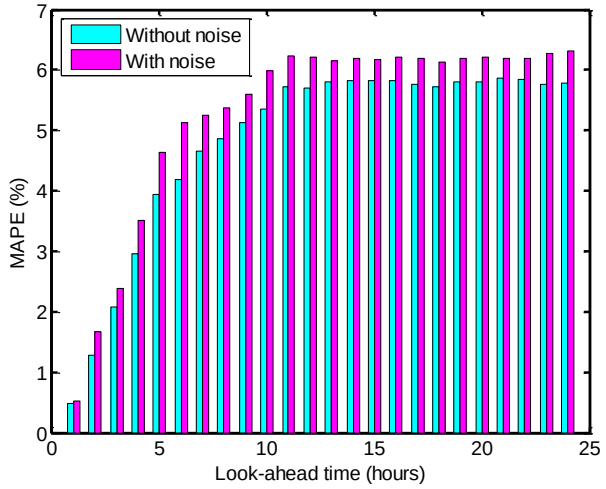


Fig. 7. 1-hour to 24-hour ahead forecasting results using measured and noisy wind speed data.

As shown in Fig. 7, the proposed method is able to achieve encouraging forecasting results with the noisy wind speed data. The average forecasting error increases about 9.22% over the whole range of 24 look-ahead hours. As the wind power series does not present long-term trends, the input variables related to wind power are always the ones with small lagged times. The quality of wind speed data, in this situation, would have a great impact in the multi-step wind power forecasting.

Case 6: The proposed method is examined using the site 50 wind farm data. It should be pointed out that the wind farm at site 50 is quite different in installed capacity and location from the one at site 1. The persistence and new reference methods in Case 1 and the LMNN, RBFNN and WNN methods in Case 2 are also involved for the purpose of comparison. The training, validation and testing periods are the same as them in Case 1. The forecasting results of 1-, 24- and 48-hour cases are shown in Table IV.

It is observed that the proposed method has obtained better forecasting performance than other models in all testing cases, which verifies its feasibility and effectiveness in WPF. Taking the 48-hour case as an example, the enhancements in MAPE of the proposed method with respect to the previous five methods are 82.1%, 75.4%, 21.8%, 23.2% and 12.3%, respectively.

TABLE IV
FORECASTING RESULTS FOR THE MODELS IN CASE 6

		1-hour	24-hour	48-hour
Persistence	MAPE	7.66	24.62	28.37
	NRMSE	11.78	31.90	36.01
New Reference	MAPE	7.91	16.43	20.62
	NRMSE	11.72	19.14	23.83
LMNN	MAPE	1.35	5.98	6.48
	NRMSE	1.92	7.63	8.28
RBFNN	MAPE	1.56	6.09	6.60
	NRMSE	2.25	7.70	8.45
WNN	MAPE	0.86	5.16	5.78
	NRMSE	1.17	6.74	7.54
Proposed	MAPE	0.44	4.44	5.07
	NRMSE	0.59	5.80	6.59

B. Discussions

Forecasting is never 100% correct, which drives us to make continuous effort to create advanced models. As clearly shown in the forecasting results, the proposed method surpasses other benchmark methods in two wind farm datasets. The improved forecasting accuracy can be attributed to several reasons, such as the proposed wavelet-based ensemble strategy, the proposed feature selection technique, the early stopping criterion and the partial least squares regression method.

The proposed method for WPF presents many advantages. First, it can tackle the difficulty induced by the nonstationarity of wind power series by means of wavelet transform. Second, it can alleviate many trivial problems in WPF, such as random network weights and biases, wavelet parameter determination, uncertainty and overtraining. Third, as shown in Case 5, it has robustness with respect to large wind speed forecasting errors. Fourth, it can produce high quality forecasting results for wind farms with different locations and sizes, which has been shown in Case 6.

IV. CONCLUSION

In this paper, a novel ensemble model for WPF is proposed, which consists of neural networks, wavelet transform, variable selection and partial least squares regression. A new ensemble scheme using different wavelet parameters is proposed to build the neural network ensemble model. A robust feature selection technique called CMIFS is developed to select the most useful input data for the forecasting model. To obtain an accurate and reliable ensemble forecast, PLSR is employed to aggregate the individual forecasts. To demonstrate the effectiveness, the proposed method has been tested on actual data from National Renewable Energy Laboratory. Compared to other forecasting methods, the proposed method is able to produce better results in the whole range of look-ahead times. The future works can be extended to probabilistic wind power forecasting and wind power ramp forecasting.

V. APPENDIX

Equation (6) describes the relationship between conditional mutual information and trivariate mutual information. It is seen

that the large value $I(C;F_j|F_i)$ could be generated by two small values $I(C;F_j)$ and $I(C;F_j;F_i)$. The new feature F_j that is weakly relevant to C can be selected in this situation. A threshold T is therefore needed to filter out the irrelevant and weakly relevant features, which is set to be 40 in this paper.

In this appendix, the influence of the predefined threshold T is studied. The constitutive component A_2 in two-level wavelet decomposition is used for experiment. The candidate set of A_2 is $\{A_2(t\text{-}1),\ldots, A_2(t\text{-}100), \mathrm{WS}(t),\ldots, \mathrm{WS}(t\text{-}100)\}$. The number of features in the subset S is 25. The selection results with and without using the threshold are shown in Table V.

TABLE V
FEATURE SELECTION RESULTS WITH AND WITHOUT USING THRESHOLD

	Selected features
Without threshold T	$A_2(t\text{-}1)$, $A_2(t\text{-}2)$, $A_2(t\text{-}3)$, $A_2(t\text{-}4)$, $A_2(t\text{-}5)$, $A_2(t\text{-}6)$, $A_2(t\text{-}7)$, $A_2(t\text{-}8)$, $A_2(t\text{-}9)$, $A_2(t\text{-}21)$, $A_2(t\text{-}29)$, $A_2(t\text{-}30)$, $A_2(t\text{-}31)$, $A_2(t\text{-}93)$, WS(t-0), WS(t-1), WS(t-2), WS(t-3), WS(t-4), WS(t-5), WS(t-6), WS(t-7), WS(t-8), WS(t-9), WS(t-10)
With threshold T	$A_2(t\text{-}1)$, $A_2(t\text{-}2)$, $A_2(t\text{-}3)$, $A_2(t\text{-}4)$, $A_2(t\text{-}5)$, $A_2(t\text{-}6)$, $A_2(t\text{-}7)$, $A_2(t\text{-}8)$, $A_2(t\text{-}9)$, $A_2(t\text{-}10)$, $A_2(t\text{-}20)$, $A_2(t\text{-}21)$, $A_2(t\text{-}22)$, WS(t-0), WS(t-1), WS(t-2), WS(t-3), WS(t-4), WS(t-5), WS(t-6), WS(t-7), WS(t-8), WS(t-9), WS(t-10), WS(t-11)

It is seen that the selection results with and without using T are different. In the candidate set, the maximum and minimum MI values are 0.9874 and 0.00014, respectively. The MI value associated with the threshold is 0.0096. In the case of without T, the MI values of $A_2(t\text{-}29)$, $A_2(t\text{-}30)$, $A_2(t\text{-}31)$ and $A_2(t\text{-}93)$ are 0.0030, 0.0021, 0.0015 and 0.00096, which are smaller than 0.0096. These four features can be regarded as weakly relevant features and should be removed. The comparison confirms that setting a threshold can avoid including irrelevant and weakly relevant features in the selected subset S.

VI. REFERENCES

[1] Global wind energy outlook 2014. Brussels, Belgium: Global Wind Energy Council, 2014.

[2] A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh, "Current methods and advances in forecasting of wind power generation," *Renew. Energy*, vol. 37, pp. 1-8, 2012.

[3] R. G. Kavasseri and K. Seetharaman, "Day-ahead wind speed forecasting using f-ARIMA models," *Renew. Energy*, vol. 34, pp. 1388-1393, 2009.

[4] H. Liu, E. Erdem, and J. Shi, "Comprehensive evaluation of ARMA-GARCH(-M) approaches for modeling the mean and volatility of wind speed," *Appl. Energy*, vol. 88, pp. 724-732, 2011.

[5] P. Louka, G. Galanis, N. Siebert, G. Kariniotakis, P. Katsafados, I. Pytharoulis, *et al.*, "Improvements in wind speed forecasts for wind power prediction purposes using Kalman filtering," *J. Wind Eng. Ind. Aerodyn.*, vol. 96, pp. 2348-2362, 2008.

[6] N. Amjady, F. Keynia, and H. Zareipour, "Wind power prediction by a new forecast engine composed of modified hybrid neural network and enhanced particle swarm optimization," *IEEE Trans. Sustain. Energy*, vol. 2, pp. 265-276, 2011.

[7] K. Bhaskar and S. N. Singh, "AWNN - assisted wind power forecasting using feed-forward neural network," *IEEE Trans. Sustain. Energy*, vol. 3, pp. 306-315, 2012.

[8] Y. Liu, J. Shi, Y. Yang, and W. J. Lee, "Short-term wind power prediction based on wavelet transform-support vector machine and statistic-characteristics analysis," *IEEE Trans. Ind. Appl.*, vol. 48, pp. 1136-1141, 2012.

[9] J. P. S. Catalao, H. M. I. Pousinho, and V. M. F. Mendes, "Hybrid wavelet-PSO-ANFIS approach for short-term wind power forecasting in Portugal," *IEEE Trans. Sustain. Energy*, vol. 2, pp. 50-59, 2011.

[10] A. Togelou, G. Sideratos, and N. D. Hatziargyriou, "Wind power forecasting in the absence of historical data," *IEEE Trans. Sustain. Energy*, vol. 3, pp. 416-421, 2012.

[11] L. Ma, S. Luan, C. Jiang, H. Liu, and Y. Zhang, "A review on the forecasting of wind speed and generated power," *Renew. Sustain. Energy Rev.*, vol. 13, pp. 915-920, 2009.

[12] I. G. Damousis, M. C. Alexiadis, J. B. Theocharis, and P. S. Dokopoulos, "A fuzzy model for wind speed prediction and power generation in wind parks using spatial correlation," *IEEE Trans.Energy Convers.*, vol. 19, pp. 352-361, 2004.

[13] N. Chen, Q. Zheng, I. T. Nabney, and X. Meng, "Wind power forecasts using Gaussian processes and numerical weather prediction," *IEEE Trans. Power Syst.*, vol. 29, pp. 656-665, 2014.

[14] J. Jung and R. P. Broadwater, "Current status and future advances for wind speed and power forecasting," *Renew. Sustain. Energy Rev.*, vol. 31, pp. 762-777, 2014.

[15] G. Brown, J. L. Wyatt, and P. Ti, "Managing diversity in regression ensembles," *J. Mach. Learn. Res.*, vol. 6, pp. 1621-1650, 2005.

[16] J. W. Taylor, P. E. McSharry, and R. Buizza, "Wind power density forecasting using ensemble predictions and time series models," *IEEE Trans.Energy Convers.*, vol. 24, pp. 775-782, 2009.

[17] L. Duehee and R. Baldick, "Short-term wind power ensemble prediction based on Gaussian processes and neural networks," *IEEE Trans. Smart Grid*, vol. 5, pp. 501-510, 2014.

[18] P. Pinson, H. A. Nielsen, H. Madsen, and G. Kariniotakis, "Skill forecasting from ensemble predictions of wind power," *Appl. Energy*, vol. 86, pp. 1326-1334, 2009.

[19] A. Khosravi and S. Nahavandi, "Combined nonparametric prediction intervals for wind power generation," *IEEE Trans. Sustain. Energy*, vol. 4, pp. 849-856, 2013.

[20] A. U. Haque, P. Mandal, J. Meng, A. K. Srivastava, T. L. Tseng, and T. Senjyu, "A novel hybrid approach based on wavelet transform and fuzzy ARTMAP networks for predicting wind farm power production," *IEEE Trans. Ind. Appl.*, vol. 49, pp. 2253-2261, 2013.

[21] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Anal. Chim. Acta.*, vol. 185, pp. 1-17, 1986.

[22] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157-1182, 2003.

[23] N. Kwak and C. H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 1667-1671, 2002.

[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken: Wiley Interscience, 2006.

[25] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205-1224, 2004.

[26] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531-1555, 2004.

[27] S. S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. New York: Prentice Hall, 2009.

[28] D. Nguyen and B. Widrow, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," in *Proc. Int. Joint Conf. on Neural Networks*, 1990, pp. 21-26.

[29] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics, 1992.

[30] A. J. R. Reis and A. P. A. da Silva, "Feature extraction via multiresolution analysis for short-term load forecasting," *IEEE Trans. Power Syst.*, vol. 20, pp. 189-198, 2005.

[31] J. J. Rubio, "Adaptive least square control in discrete time of robotic arms," *Soft Comput.*, pp. 1-12, 2014.

[32] G. Özkan and M. İnal, "Comparison of neural network application for fuzzy and ANFIS approaches for multi-criteria decision making problems," *Appl. Soft Comput.*, vol. 24, pp. 232-238, 2014.

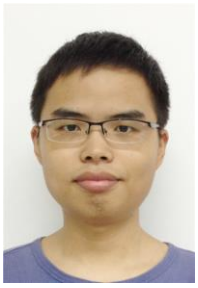
[33] S. de Jong, "SIMPLS: an alternative approach to partial least squares regression," *Chemometr. Intell. Lab. Syst.*, vol. 18, pp. 251-263, 1993.

[34] National Renewable Energy Laboratory [Online]. Available: http://www.nrel.gov/electricity/transmission/eastern_wind_methodology.html

[35] J. J. Rubio, "Stable and optimal controls of a proton exchange membrane fuel cell," *Int. J. of Control*, vol. 87, pp. 2338-2347, 2014.

[36] M. Sheikholeslami, F. B. Sheykholeslami, S. Khoshhal, H. Mola-Abasia, D. D. Ganji, and H. B. Rokni, "Effect of magnetic field on Cu-water nanofluid heat transfer using GMDH-type neural network," *Neural Comput. Appl.*, vol. 25, pp. 171-178, 2014.
[37] S. S. Soman, H. Zareipour, O. Malik, and P. Mandal, "A review of wind power and wind speed forecasting methods with different time horizons," in *North American Power Symposium*, 2010, pp. 1-8.
[38] H. Madsen, P. Pinson, G. Kariniotakis, H. A. Nielsen, and T. Nielsen, "Standardizing the performance evaluation of shor-term wind power prediction models," *Wind Eng.*, vol. 29, pp. 475-489, 2005.
[39] R. S. Marko and K. Igor, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, pp. 23-69, 2003.
[40] H. Peng, L. Fulmi, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 1226-1238, 2005.

VII. BIOGRAPHIES



Song Li received the B.Eng degree from University of Electronic Science and Technology of China, Chengdu, China, in 2010. Currently he is pursuing the PhD degree in School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

Peng Wang (M'00–SM'11) received the B.Sc. degree from Xi'an Jiaotong University, Xian, China, in 1978; the M.Sc. degree from the Taiyuan University of Technology, Taiyuan, China, in 1987; and the M.Sc. and Ph.D. degrees with specialization in power engineering from the University of Saskatchewan, Saskatoon, SK, Canada, in 1995 and 1998, respectively.

He is currently an Associate Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

Lalit Goel (F'13) received the B.Tech. degree in electrical engineering from the Regional Engineering College, Warangal, India, in 1983 and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Saskatchewan, Saskatoon, Canada, in 1988 and 1991 respectively.

He joined the School of Electrical and Electronic Engineering (EEE) at the Nanyang Technological University (NTU), Singapore, in 1991, where he is currently a Professor of Power Engineering and Director of the Office of Global Education & Mobility.