Forecasting the Wind Generation Using a Two-Stage Network Based on Meteorological Information

Shu Fan, *Member, IEEE*, James R. Liao, *Member, IEEE*, Ryuichi Yokoyama, *Senior Member, IEEE*, Luonan Chen, *Senior Member, IEEE*, and Wei-Jen Lee, *Fellow, IEEE*

***Abstract*—This paper proposes a practical and effective model for the generation forecasting of a wind farm with an emphasis on its scheduling and trading in a wholesale electricity market. A novel forecasting model is developed based on indepth investigations of meteorological information. This model adopts a two-stage hybrid network with Bayesian clustering by dynamics and support vector regression. The proposed structure is robust with different input data types and can deal with the nonstationarity of wind speed and generation series well. Once the network is trained, we can straightforward predict the 48-h ahead wind power generation. To demonstrate the effectiveness, the model is applied and tested on a 74-MW wind farm located in the southwest Oklahoma of the United States.**

***Index Terms*—Machine learning, meteorology, nonstationarity, wind generation forecasting.**

I. Introduction

WIND POWER is becoming the fastest growing and a mature renewable energy source in the world. By the end of 2006, global wind power nameplate capacity has exceeded 74 200 MW, which represents nearly 26% of increase in merely a year's time. [1] It is estimated that by 2020, about 12% of the world's electricity will be supplied by wind generation.

Although the integration of wind power brings significant environmental and economical benefits, the intermittent and stochastic nature of wind energy also presents challenges in power system operating and planning. Because a power system must maintain instantaneous balancing between the aggregated generation and demand at all times, variations in wind farm output will increase the regulation requirements and reduce the operational efficiency of some generating units.

A possible solution to such a problem is to improve the wind generation forecast. However, the accurate forecasting of wind generation is a very demanding task due to highly complex interactions and contributions of various parameters to the forecasting.

In the past years, many approaches have been proposed for the wind generation forecasting. The particular forecasting method used depends on the available information and the time scale of the application. Basically, the forecasting from milliseconds to seconds is used for wind turbines control [2], [3], whereas, the time scales from minutes to hours or even weeks are important for the integration of wind power into the energy system [4]–[10].

This paper focuses on the hourly wind generation forecasts for a wind farm with the time scale given by the electricity market, from 1 to 48 h ahead. For such a forecast horizon, besides historical wind speed and generation data, the weather forecasts from a numerical weather prediction (NWP) model or meteorological services provide the essential information for the forecast. Two basic principles are used to forecast the wind power: physical models and statistical approaches [4]. The physical models use physical considerations to reach the best possible estimate of the local wind speed before using model output statistics (MOS) to reduce the remaining error. Statistical models find the relationships between a set of explanatory variables including NWP results and online measured generation data. The model to be adopted usually depends on the specific application and availability of information.

The purpose of this paper is to propose a practical and cost-efficient model for the generation forecasting of a wind farm. To produce the forecast with a high degree of accuracy and a minimum amount of effort to modify the input data for different wind farms, a statistical forecasting model has been selected for the research in this paper.

To establish a statistical forecasting model, two key problems are required.

1) *Nonstationarity of Time Series*. The wind generation (WG)/wind speed (WS) time series exhibit strong nonstationarity due to the multiple seasonality and a high percentage of abrupt changes. In other words, the WG/WS series switch between different regimes, which generally give rise to piecewise-stationary dynamics. However, almost all the methods of time series analysis, both linear and nonlinear, assume some stationarity of the system under investigation. Therefore, how to handle the nonstationarity is a key challenge in modeling time series of wind power.
2) *Robustness of the Forecasting System*. Many exogenous variables besides the WG/WS, such as the wind direction, elevation, atmospheric pressure, temperature and humidity, etc., should be taken into account in the forecasting. These variables are usually dealt with in a unity model,

Manuscript received July 12, 2006; revised September 10, 2007. First published January 20, 2009; current version published May 19, 2009. Paper no. TEC-00248-2006.

S. Fan is with the Business and Economic Forecasting Unit, Monash University, Clayton, Vic. 3800, Australia (e-mail: shu.fan@buseco.monash.edu.au).

J. R. Liao is with the Western Farmers Electric Cooperative, Anadarko, OK 73005 USA (e-mail: j_liao@wfec.com).

R. Yokoyama is with the Waseda University, Tokyo 169-8555, Japan (e-mail: yokoyama-ryuichi@waseda.jp).

L. Chen is with Osaka Sangyo University, Osaka 574-0013, Japan (e-mail: chen@eic.osaka-sandai.ac.jp).

W.-J. Lee is with the Energy Systems Research Center, University of Texas, Arlington, TX 76013 USA (e-mail: wlee@uta.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.org.

Digital Object Identifier 10.1109/TEC.2008.2001457

0885-8969/$25.00 © 2009 IEEE

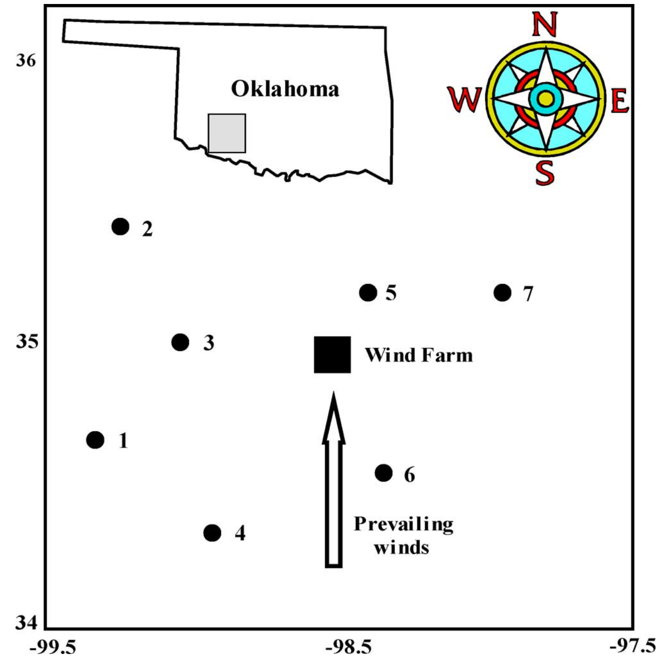

Fig. 1. Geographical location of blue canyon wind farm with altitude and longitude labeled in the map.

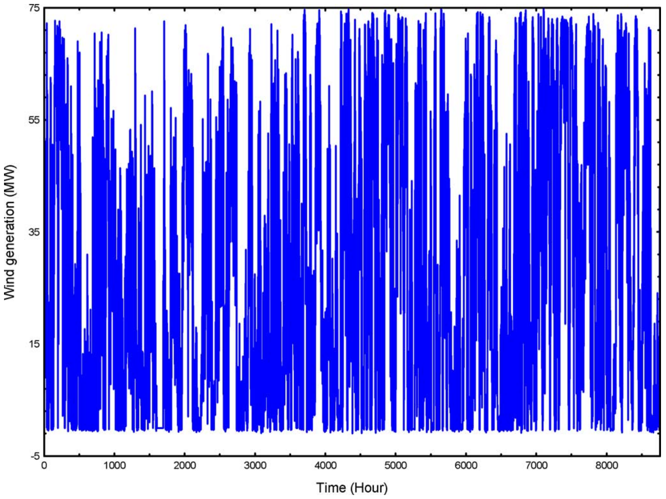

Fig. 2. Ten-minutes wind generation in the wind farm from June 1, 2004 to May 31, 2005.

such as a neural network with multiple inputs and outputs. This usually makes the forecasting tools variable-sensitive and system-dependent. A model developed for one wind farm cannot be easily modified for another, which decreases the robustness of the forecasting model.

Based on the aforementioned analysis, a novel and effective forecasting model with Bayesian clustering by dynamics (BCD) and support vector regression (SVR) is proposed in this paper with emphasis on tackling the two key problems. The proposed model is well suited for capturing the dynamics of WG/WS time series by using hybrid architecture, and it has strong robustness and can be easily modified for different wind farms. Moreover, it is established on meteorological information that is easy to access. Experiments and comparisons with the persistence method demonstrate the effectiveness of the proposed model in learning and predicting wind generation.

II. Task Description and Data Analysis

This paper uses the Blue Canyon I wind farm, located in southwestern Oklahoma, United States, as the testing system for our method. The wind farm began commercial operation in December 2003 with a nameplate capacity of 74 MW. Its geographical location is shown in Fig. 1.

The dots in Fig. 1 refer to the weather stations within a radius of about 80 km of the wind farm center. There are two meteorological towers within the wind farm where 45 wind turbines distributed seven miles east–west and two miles north–south are placed directly perpendicular to the prevailing wind direction.

The following data have been collected for the model establishment.

1) Ten-minutes data from the wind farm, including the observations of aggregated wind generation of the wind farm and the observations of wind speed and direction from a meteorological tower within the farm. In this paper, the 10-minutes data were averaged in 1-h intervals.
2) Hourly observations of weather data from the surrounding weather stations, including the wind speed and direction, temperature, relative humidity, and atmospheric pressure.
3) Hourly meteorological forecasts at the locations of wind farm and the weather stations. These near-surface node predictions are obtained from the commercial weather services using the NWP model. The forecasts are formulated as records comprising hourly predictions of the wind speed and direction and the other weather data at the succeeding 48 h ahead.

Fig. 2 illustrates the aggregated wind generation of the wind farm from June 1, 2004 to May 31, 2005. According to Fig. 2, it is clear that the wind generation curve is widely divergent and fluctuates with a high frequency, and there are also a high percentage of abrupt changes or spikes in the curve, indicating nonstationarity of the time series.

To give a quantitative index for the nonstationarity of the wind generation time series in Fig. 2, we have carried out an RQA analysis here. RQA is the extension of recurrence plot analysis [11], which is designed to locate hidden recurring patterns, nonstationarity, and structural changes in time series. The *Trend* statistic in RQA can provide information about nonstationarity in a process, especially a drift. Nonzero trend indicates drift in the system, while zero (or very close to zero) values indicate stationarity [12]. In this paper, we applied this analysis to the wind generation series in Fig. 2. RQA was performed with embedding dimension equal to 3, and the computations were repeatedly performed on the time series data within episodic windows consisting of 144 consecutive points. Sequential windows were shifted by 30 points, granting a total of 1747 trend values. The results are plotted in Fig. 3.

As shown in Fig. 3, the trend statistic departs away from zero most of the time, and there exist many negative spikes,

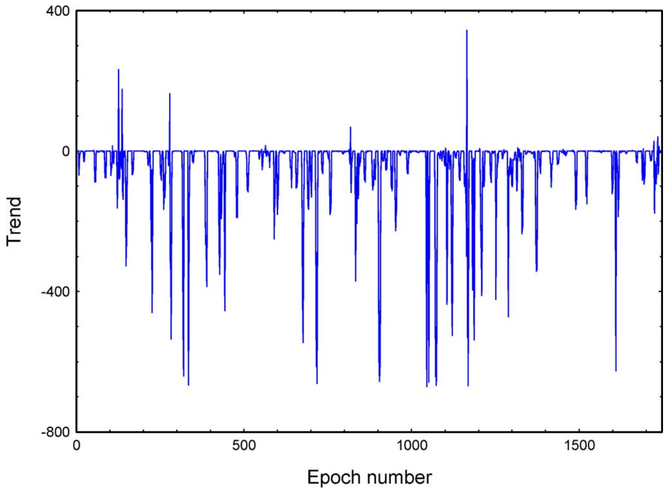

Fig. 3. Trend statistic for ten-minutes wind generation series. RAQ parameters: delay = 6; embedding dimension = 3; radius = 1.5; line definition = 10 points.

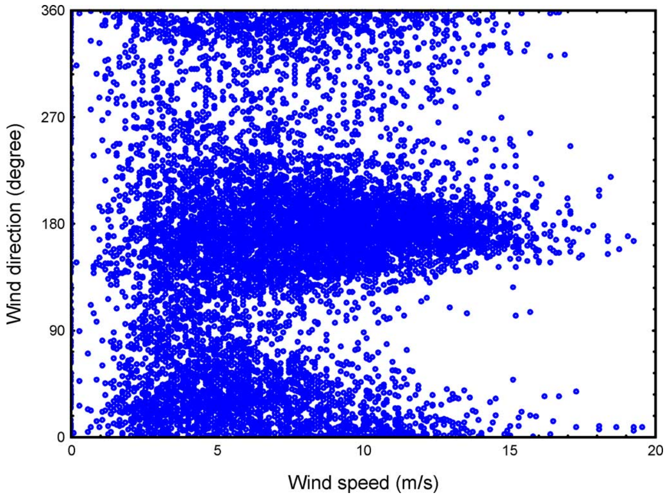

Fig. 4. Correlation between the local wind speed and the direction from June 1, 2004 to May 30, 2005. 0 or 360 is recorded as north, 90 = east, 180 = south, and 270 = west.

which demonstrate the existence of nonstationarity in the wind generation series.

Furthermore, Fig. 4 shows the relationship between the wind speed and direction. From this figure, we can find small winds of all directions; however, the directions are narrowed on the south and northeast band. This phenomenon further demonstrates that different patterns exist in the wind speed series.

It is clear that the wind speed is the key information for the generation forecasting. So, it is necessary to analyze the correlation between the aggregated wind generation and the local wind speed within the wind farm. Fig. 5 shows this correlation from June 1, 2004 to May 30, 2005, which can be computed using the following expression:

$$\rho_{g,s} = \frac{\mathrm{Cov}(g,s)}{\sigma_g \sigma_s} \tag{1}$$

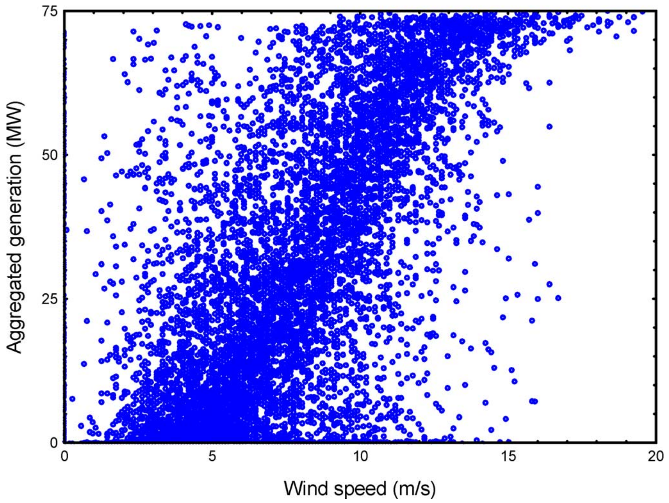

Fig. 5. Correlation between the wind speed and the aggregated wind generation at the wind farm from June 1, 2004 to May 30, 2005.

where $\mathrm{Cov}(g,s)$ is the covariance of generation g and speed s, and σ_g and σ_s are the standard deviations for g and s.

The historical correlation for data in Fig. 5 is 0.66, indicating a close relationship between the two variables. On the other hand, it is also obvious that the wind power may vary widely at the same measured wind speed of the local site, mainly because the wind turbines are distributed along a wide geographical area, indicating that spatial correlation exists between the wind generation and the wind speed. Therefore, it is necessary to use the data from both local and remote sites in the forecasting.

Based on the earlier analysis, the forecasting system will be established based on the meteorological information in various nodes around the wind farm, taking advantage of possible spatial correlation of wind time series between the wind farm and the surrounding positions.

III. Method and the Learning Algorithm

A. Architecture of the Forecasting System

A time-series-based nonlinear discrete-time dynamical model, which is represented by (2), will be applied to the forecasting

$$y(t+1) = f(y(t), \ldots, y(t-m+1); X) \tag{2}$$

where $y(t)$ represents the hourly aggregated wind generation of the wind farm at time t and m is the order of the dynamical system, which is a predetermined constant. X is a vector representing the control parameters of the dynamical system, including wind speed, wind direction, and humidity. Given the meteorological predictions, the task is to extrapolate past wind generation behavior while taking into account the other influencing factors.

The proposed forecasting system is shown in Fig. 6. This model is based on hybrid architecture. First, a BCD classifier is applied to cluster the input training dataset into several subsets with similar dynamical properties in an unsupervised manner. Then, a group of SVRs was used to fit the training data in

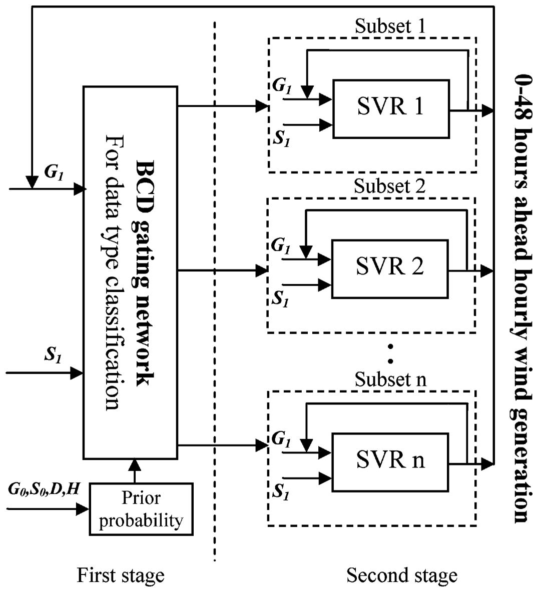

Fig. 6. Hybrid model for the wind generation forecasting.

each subset in a supervised way. In other words, depending on their stationarity as well as other dynamic features, this method breaks the time series into different segments or subsets where data in the same segment can be modeled by the same SVR due to similar property.

This hybrid structure is well suited for capturing the dynamics of wind generation time series. BCD is based on unsupervised learning, which has the ability to partition the space of input training dataset into many subsets without prior knowledge about the classifying criteria [13]. Compared to other clustering methods such as hierarchical clustering or self organizing maps (SOM), BCD identifies the set of clusters with maximum posterior probability without requiring any prior input about the number of clusters, thereby avoiding the risk of overfitting. SVR is a new and powerful machine learning technique for data regression based on recent advances in statistical learning theory [15]. Established on the unique theory of the structure risk minimization principle to estimate a function by minimizing an upper bound of the generalization error, SVRs are shown to be very resistant to the overfitting problem, eventually achieving a high generalization performance in solving forecasting problems of various time series [16], [17].

In Fig. 6, the input variables are different for the BCD and SVR networks. All the input variables are decomposed into two groups: only wind generation and speed data are used for the SVRs; for BCD, besides WG/WS data, other factors such as the wind direction and humidity will be used to capture the dynamics. This arrangement alleviates the sensitivity of the model to the variables without losing useful information. And the user can intentionally choose appropriate input variables for the BCD classifier using his knowledge and experience.

The proposed model is trained as a one-step-ahead forecast predictor. After the parameters of the model are determined in the training procedure, the wind generation forecasting can be conducted by the trained network in a voting manner among the BCD and SVR, where the output of only one SVR model is used for the final forecast. When forecasting the 48-h-ahead wind generation, some of the actual WG values are unknown in the forecasting procedure, the network's predictions at previous time steps will then be used. Consequently, (2) should be rewritten as (3), which means the 48-h estimates are recurrently derived by using past and forecast values of the inputs, as well as the network's outputs at previous time steps:

$$\hat{y}(t+1) = f(\hat{y}(t), \hat{y}(t-1), \ldots, X) \quad (3)$$

where $\hat{y}(t+1)$ represents the one-step-ahead wind generation forecast, $\hat{y}(t), \hat{y}(t-1), \ldots$ represent predictions of previous time steps, and X is same as that in (2).

TABLE I
CROSS-CORRELATIONS BETWEEN AGGREGATED WIND GENERATION AND DIFFERENT OBSERVATIONS WITHIN THE WIND FARM

	Wind direction	Barometric pressure	Temperature	Humidity	Gust speed
Aggregated WG	0.14	-0.02	-0.02	-0.07	0.62

B. Selection of Input Variables

Selection of the model's inputs variables has a great effect on the performance of the forecasting models, and should be properly addressed. The following aspects have to be considered in selecting the input variables for the model. First, although many exogenous variables have relationship with the wind generation, it is neither feasible nor efficient to use all of them in the forecasting. Hence, we should pick the ones exhibiting a significant degree of correlation with regard to the wind generation. Second, as indicated in the previous section, seven candidate weather stations are available for providing the meteorological data, so it is necessary for us to determine whether the data from a specific station are suitable for the forecasting or not. Finally, we need to know how many lagged hours should be included in the WG/WS time series. These selections can be achieved by exploiting the analysis of autocorrelation and cross-correlation between the different variables.

Table I shows the cross-correlations between aggregated wind generation and different observations within the wind farm for the period in Fig. 2. At the same time, we have also investigated these cross-correlations in different months. It can be concluded that the wind generation also has a relatively close relationship with wind direction and gust speed. However, the correlation coefficients are not consistent and usually shift over time. As for the barometric pressure, temperature, and humidity, their influences on the generation are not evident.

Table II gives the cross-correlation coefficients between the wind speed in different stations and the aggregated generation of the wind farm for the period in Fig. 2. We also conducted correlation analysis for each season of the year, and the results are basically consistent with those in Table II. Since the wind directions in different weather stations are generally similar during certain period, the correlations in Table II are used to select the data of weather stations. It can be seen that the coefficients

TABLE II
CROSS-CORRELATIONS BETWEEN THE WIND SPEEDS IN THE WEATHER STATIONS AND THE AGGREGATED GENERATION OF THE WIND FARM

Stations	1	2	3	4	5	6	7
Cross-correlation With generation	0.56	0.66	0.67	0.54	0.62	0.50	0.51

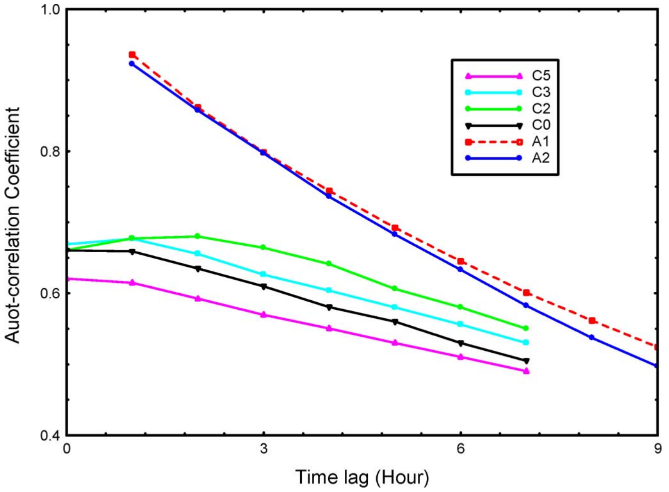

Fig. 7. Auto-correlation (A) curves for aggregated wind generation and speed within the wind farm and cross-correlation (C) curves for aggregated wind generation and speed in different locations.

TABLE III
LIST OF INPUT DATA OF THE SVR NETWORK

Input	Variable name	Lagged value (hours)
1-6	Hourly wind power generation (G_1)	1,2,3,4,5,6
7-12	Hourly wind speed 0 (S_1)	0,1,2,3,4,5
13-18	Hourly wind speed 2	0,1,2,3,4,5
19-24	Hourly wind speed 3	0,1,2,3,4,5
25-30	Hourly wind speed 5	0,1,2,3,4,5

The wind speed 0 means the speed within the wind farm, and accordingly number 2, 3, or 5 means weather station 2, 3, or 5.
The hour of the predication is assumed at 0, the lag 0 represents the target instant, and the 6 lagged hours means the values that were measured 6 h earlier than the hour of predication.

for stations 1, 4, 6, and 7 are below 0.60, so the data from these stations will not be used in the model.

Fig. 7 illustrates autocorrelation (A) curves for aggregated wind generation and speed within the wind farm and cross-correlation (C) curves for aggregated wind generation and speed in different locations.

Based on the previous data analysis, the input variables for BCD and SVR can then be selected in Tables III and IV.

As indicated in Table III, the input data of SVR consist of the following elements: hourly wind generation series of the previous hours and the forecasted and observed wind speed series of the wind farm and the weather stations. In order to capture the time series style, we include the wind generations of previous 6 h. Besides the forecasted wind speed at the target hour, the wind speeds of previous 5 h are also included.

The input data of the BCD network are shown in Table IV. In addition to the time series G_1 and S_1 as the primary input for clustering the hour by dynamics, the other factors, including average wind generation of previous 6 h, average wind speed in different locations, the forecasted average wind direction and humidity of the target hour, are used to determine the prior probability for the BCD. Other factors can also be included, and the selection of input data mainly depends on the system to be studied.

TABLE IV
LIST OF INPUT DATA OF THE BCD CLASSIFIER

Input	Variable	Detail description
1-6	G_1	Wind power generation series
7-30	S_1	Wind speed series
31-40	Variables to determine prior probability	G_0: Average wind generation
		S_0: Average wind speed
		D: Forecasted wind direction
		H: Average humidity

In the experimentation, we found that inclusion of the temperature and the atmospheric pressure as inputs did not offer improved performance of the resulting models. On the contrary, the presence of two more inputs was actually slowing down the learning process. Therefore, these variables are also not included in the input set.

C. *Learning Algorithm: The BCD Classifier*

The clustering method implemented in BCD is based on a novel concept of similarity for time series: two time series are similar when they are generated by the same stochastic process. BCD models time series by autoregressive equations [13]. Let $S_j = \{x_{j1}, \ldots, x_{jt}, \ldots, x_{jn}\}$ denote a stationary time series of continuous values. The series follows an autoregressive model of order p, say AR(p), if the value of the series at time $t > p$ is a linear function of the values observed in the previous p steps. We can describe the model in a matrix form as

$$x_j = X_j \beta_j + \varepsilon_j \qquad (4)$$

where x_j is the vector $(x_{j(p+1)}, \ldots, x_{jn})^T, X_j$ is the $(n-p) \times (p+1)$ regression matrix whose t-th row is $(1, x_{j(t-1)}, \ldots, x_{j(t-p)})$ for $t > p, \beta_j$ is the vector of autoregressive coefficients, and ε_j is the vector of uncorrelated errors that are assumed normally distributed with expected value $E(\varepsilon_{jt}) = 0$ and variance $V(\varepsilon_{jt}) = \sigma_j^2$, for any time point t. Given the data, the model parameters can be estimated using standard Bayesian procedures, and details are described in [14].

To select the set of clusters, BCD uses a novel model-based Bayesian clustering procedure. A set of clusters $C_1, \ldots, C_k, \ldots, C_c$, each consisting of m_k time series, is represented as a model M_C.

The time series assigned to each cluster are treated as independent realizations of the dynamic process represented by the cluster, which is described by an autoregressive equation. Each cluster C_k can be jointly modeled as

$$x_k = X_k \beta_k + \varepsilon_k$$

where the vector x_k and the matrix X_k are defined by stacking the m_k vectors x_{kj} and regression matrices X_{kj}, one for each time series, as follows:

$$x_k = \begin{pmatrix} x_{k1} \\ \vdots \\ x_{km_k} \end{pmatrix}, \qquad X_k = \begin{pmatrix} X_{k1} \\ \vdots \\ X_{km_k} \end{pmatrix}.$$

Given a set of possible clustering models, the task is to rank them according to their posterior probability. The posterior probability of the model M_C is computed by Bayes theorem as

$$P(M_C \mid x) \propto P(M_C) f(x \mid M_C) \tag{5}$$

where $P(M_C)$ is the prior probability of M_C and $f(x \mid M_C)$ is the marginal likelihood. Assuming independent uniform prior distributions on the model parameters and a symmetric Dirichlet distribution on the cluster probability p_k, the marginal likelihood of each cluster model M_C can be easily computed in a closed form by solving the integral

$$f(x \mid M_C) = \int f(x \mid \theta_C) f(\theta_C)\, d\theta_C \tag{6}$$

where θ_C is the vector of parameters that describe the likelihood function, conditional on a clustering model M_C, and $f(\theta_C)$ is the prior density. In this way, each clustering model has an explicit probabilistic score and the model with maximum score can be found. In particular, $f(x \mid M_C)$ can be computed as

$$\begin{aligned} f(x \mid M_c) &= \frac{\Gamma(1)}{\Gamma(1+m)} \times \prod_{k=1}^{c} \frac{\Gamma(m_k/m + m_k)}{\Gamma(m_k/m)} \\ &\quad \times \frac{(\mathrm{RSS}_k/2)^{(q-n_k)/2}\, \Gamma\left(n_k - q\right)/2}{(2\pi)^{(q-n_k)/2} \det(X_k^T X_k)^{(1/2)}} \end{aligned} \tag{7}$$

where n_k is the dimension of the vector x_k and $\mathrm{RSS}_k = x_k^T(I_n - X_k(X_k^T X_k)^{-1} X_k^T)x_k$ is the residual sum of squares in cluster C_k. When all clustering models are *a priori* equally likely, the posterior probability $P(M_C \mid x)$ is proportional to the marginal likelihood $f(x \mid M_C)$, which becomes our probabilistic scoring metric.

As the number of clusters or subsets grows exponentially with the number of time series, BCD uses an agglomerative search strategy, which iteratively merges time series into clusters. The procedure starts by assuming that each of the m wind generation time series is generated by a different process. Thus, the initial model M_m consists of m clusters, one for each time series, with score $f(x \mid M_m)$. The next step is the computation of the marginal likelihood of the $m(m-1)$ models in which two of the m profiles are merged into one cluster. The model M_{m-1} with maximal marginal likelihood is chosen and the merging is rejected if $f(x \mid M_m) \geq f(x \mid M_{m-1})$ and the procedure stops. If $f(x \mid M_m) < f(x \mid M_{m-1})$, the merging is accepted and a cluster C_k merging the two time series is created. In such a way, the procedure is repeated on the new set of $m-1$ time series that consist of the remaining $m-2$ time series and the cluster profile.

Although the agglomerative strategy makes the search process feasible, the computational effort can be extremely demanding when the number of time series is large. To further reduce this effort, a heuristic strategy based on a measure of similarity between time series is applied. The intuition behind this strategy is that the merging of two similar time series has better chances of increasing the marginal likelihood. The heuristic search starts by computing the $m(m-1)$ pairwise similarity measures of the time series and selects the model M_{m-1} in which the two closest time series are merged into one cluster. If$f(x \mid M_m) < f(x \mid M_{m-1})$, the two time series are merged into a single cluster, a profile of this cluster is computed by averaging the two observed time series, and the procedure is repeated on the new set of $m-1$ time series. If this merging is rejected, the procedure is repeated on the two time series with the second highest similarity until an acceptable merging is found. If no acceptable merging is found, the procedure stops. Note that the clustering procedure is actually performed on the posterior probability of the model and the similarity measure is only used to increase the speed of the search process and to limit the risk of falling into local maxima.

Similarity measures of two time series implemented in BCD include Euclidean distance correlation, and Kullback–Leiber distance. In numerical experiments, we have tried different distances and finally adopted the Euclidean distance of two time series $S_j = \{x_{i1}, \ldots, x_{in}\}$ and $S_j = \{x_{j1}, \ldots, x_{jn}\}$, computed as

$$D(S_i, S_j) = \sqrt{\sum_{t=1}^{n} (x_{it} - x_{jt})^2}. \tag{8}$$

D. Learning Algorithm: The SVR Network

Suppose that we are given training data $(x_1, y_1), \ldots (x_i, y_i), \ldots (x_n, y_n)$, where x_i are input patterns and y_i are the associated output value of x_i, the support vector regression solves an optimization problem [15]

$$\begin{aligned} &\min_{\omega, b, \xi, \xi^*} \frac{1}{2}\omega^T \omega + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ &\text{subject to}\, y_i - (\omega^T \phi(x_i) + b) \leq \varepsilon + \xi_i^* \\ &\qquad\qquad (\omega^T \phi(x_i) + b) - y_i \leq \varepsilon + \xi_i \\ &\qquad\qquad \xi_i, \xi_i^* \geq 0, \quad i = 1, \ldots, n \end{aligned} \tag{9}$$

where x_i is mapped to a higher dimensional space by the function Φ, and ξ_i^* is slack variables of the upper training error (ξ_i is the lower) subject to the ε-insensitive tube $(\omega^T \phi(x_i) + b) - y_i \leq \varepsilon$. The constant $C > 0$ determines the tradeoff between the flatness and losses. The parameters that control regression quality are the cost of error C, the width of the tube ε, and the mapping function Φ.

The constraints of (9) imply that we put most data x_i in the tube ε. If x_i is not in the tube, there is an error ξ_i or ξ_i^* that we tend to minimize in the objective function. SVR avoids underfitting and overfitting of the training data by minimizing the training error $C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$ as well as the regularization term $\omega^T \omega/2$. For traditional least-square regression, ε is always zero and data are not mapped into higher dimensional spaces. Hence, SVR is a more general and flexible treatment on regression problems.

Since Φ might map x_i to a high- or infinite-dimensional space, instead of solving ω for (9) in a high dimension, we deal with its dual problem, which can be derived using the Lagrange theory:

$$\max_{\alpha_i,\alpha_i^*} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)^T Q(\alpha_j-\alpha_j^*) - \varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)$$

$$\text{subject to, } \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0$$

$$0 \le \alpha_i, \quad \alpha_i^* \le C, \quad i=1,\ldots,n \tag{10}$$

where $Q_{ij}=\phi(x_i)^T\phi(x_j)$, and α_i and α_i^* are the Lagrange multipliers. However, this inner product may be expensive to compute because $\phi(x)$ has too many elements. Hence, we apply "kernel trick" to do the mapping implicitly. That is, to employ some special forms, inner products in a higher space yet can be calculated in the original space. Typical examples for the kernel functions are polynomial kernel $\phi(x_i)^T\phi(x_j)=(\gamma x_1^T x_2+c_0)^d$ and RBF kernel $\phi(x_i)^T\phi(x_j)=e^{-\gamma(x_1-x_2)^2}$. Here, γ, c_0, and d are kernel parameters. They are inner products in a very high-dimensional space (or infinite-dimensional space) but can be computed efficiently by the kernel trick even without knowing $\phi(x)$.

For numerical experiments in this paper, we use the software LIBSVM [18], which is a library for support vector machines with an efficient implementation of solving (10).

IV. Numerical Experiments

A. Data Collection and Implementation

Two months have been selected to forecast and validate the performance of the proposed model, corresponding to June and December 2005. The data used to train the model are from June 1, 2004 to May 30, 2005. The test sets are completely separate from the training sets and are not used during the learning procedure. All the input variables are all scaled respectively in our program.

Upon the request of the power company running the wind farm, the 48-h forecasts are made in each morning for generation schedule and electricity markets of the next operation day.

The learning procedure of the proposed architecture is outlined as follows.

1) Classify the entire training dataset into two groups: a training set used for updating the network parameters and a verification set for testing the performance.
2) Determine prior probability of the BCD classifier according to the control variables.
3) Classify the input data type using BCD according to the dynamics of WG/WS time series.
4) In each subset of the input space, train the SVR to fit the data subset according to (10).
5) Forecast the wind generation using the dataset for verification and calculate the MAPE.
6) Tune the parameters of the model and repeat steps 1)–5) until satisfactory results are obtained.
7) Select the network parameters at the minimum of the MAPE as the final ones.

For different systems, the main difference may only lie in the clustering results. Because the BCD classifier can find both the best number of clusters and the optimal assignment of time series to clusters, the proposed model can be easily applied to different power systems.

After the training procedure is finished, the following test process is applied to verify the proposed model.

1) Identify the type of the input test data according to the information of the test hour and previous hours, using the BCD classifier.
2) Use the corresponding SVR network to output the next hour wind generation.
3) Recurrently derive the 2–48 h estimates using past values of the inputs, as well as the network's outputs at previous time steps.

B. Bayesian Clustering Analysis

In this section, we will briefly describe the clustering results of BCD and carry out RQA analysis to the WG series in each cluster. This analysis not only assists the parameter adjustment in the training procedure, but also explains the effectiveness of Bayesian clustering. Here, we will only present the analysis for the first testing dataset in the numerical experiment, since the results from the other datasets are similar.

In the numerical experiment, the whole training dataset has been partitioned into nine clusters by the BCD classifier. Then, the test data are assigned to four of them. For these four subsets, the RQA analysis is used to investigate the nonstationarity of the WG series. The RQA parameters are the same as before. The trend statistics for the WG series in the subsets are closer to zero than that in the whole dataset, indicating the effect of clustering for alleviating nonstationarity of the WG series. For simplicity, we did not plot the trend statistic in every cluster.

C. Numerical Results

The criteria to compare the performance are the normalized mean absolute error (MAE) and root mean square error (RMSE) (% of the nameplate capacity) in this paper, which indicate the accuracy of recall.

MAE is defined as

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}(|y_{ai}-y_{fi}|)/P_N \times 100\% \tag{11}$$

where y_{ai} is the actual value, y_{fi} is the forecast value, P_N is the nameplate capacity, and n is the total number of value predicted.

RMSE is given as follows:

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(y_{ai}-y_{fi})}{P_N}\right)^2} \times 100\% \tag{12}$$

TABLE V
FORECAST RESULTS FOR DIFFERENT TIME HORIZON WITH DIFFERENT INPUTS

(a) Hour ahead predication
Persistence error: MAE=7.84 %P_nRMSE=11.93 %P_n

SVR Inputs (stations)				Errors(% of P_n)		Improvement (%)	
0	3	2	5	MAE	RMSE	MAE	RMSE
O	×	×	×	7.12	11.13	9.18	6.71
O	O	×	×	6.95	10.87	11.35	8.89
O	O	O	×	6.74	10.64	14.03	10.81
O	O	O	O	6.65	10.54	15.23	11.66

(b) 24-hour ahead predication
Persistence error: MAE=21.24 %P_n RMSE=29.84 %P_n

SVR Inputs (stations)				Errors(% of P_n)		Improvement (%)	
0	3	2	5	MAE	RMSE	MAE	RMSE
O	×	×	×	15.13	21.06	28.77	29.42
O	O	×	×	14.88	20.55	29.94	31.13
O	O	O	×	14.56	20.08	31.45	32.71
O	O	O	O	14.38	19.74	32.28	33.84

(c) 48-hour ahead predication
Persistence error: MAE=25.42 %P_n RMSE=34.81 %P_n

SVR Inputs (stations)				Errors(% of P_n)		Improvement (%)	
0	3	2	5	MAE	RMSE	MAE	RMSE
O	×	×	×	16.33	22.17	35.76	36.31
O	O	×	×	16.09	21.84	36.70	37.26
O	O	O	×	15.89	21.56	37.49	38.06
O	O	O	O	15.73	21.24	38.12	38.98

For comparative study, numerical simulations comparing with Persistent (or Naive) forecast, which uses the most recent information available, are also conducted. This approach assumes that the forecast value of the wind power at the ith future time steps, $\hat{P}_{per}(t+i/t)$, is the last measured one available, $P(t)$. The persistent forecast is widely used as a benchmark for comparison [4]. The benefit gained by using the proposed model is measured as the accuracy improvement over the persistent model:

$$\mathrm{Im}\,p = (\mathrm{Err}_P - \mathrm{Err}_m)/\mathrm{Err}_p \times 100\% \tag{13}$$

where Err_p is the evaluation criterion (i.e., MAE or RMSE) of the persistence and Err_m is the evaluation criterion of the proposed model.

Table V shows the numerical results of the proposed model and the improvements over the persistent model. The forecasting errors of persistence are given above the tables. The results for several experimental cases, corresponding to different time horizons and input variables, are provided.

From Table V, the following conclusions can be derived: first, the proposed forecasting model outperforms the persistence forecast in all the situations, even in the first time step; second, the addition of data from remote stations improves forecasting accuracy, and their importance grows as the predication window extends to the future.

Fig. 8 depicts the MAE and RMSE performance obtained by the proposed model, along with the performance obtained by persistence. The percentage improvement over the persistence obtained by the proposed model is illustrated in Fig. 9.

As shown in the two figures, the proposed model can provide much more accurate forecast in the whole forecast horizon; it is able to produce robust multistep ahead estimations compared

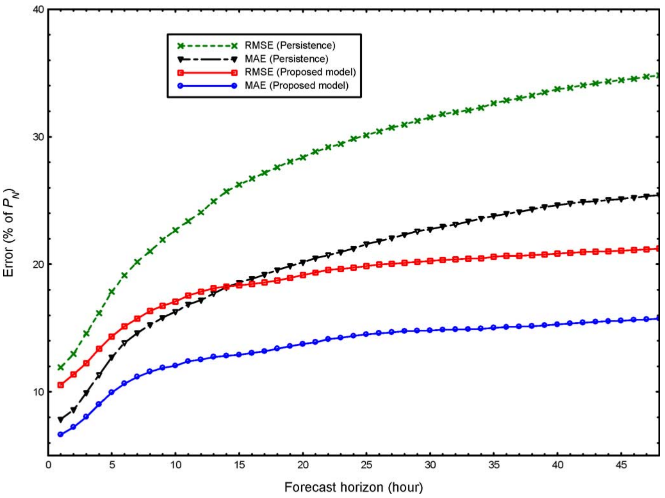

Fig. 8. Comparisons of forecast errors for a horizon of 48 h between the proposed model and persistence model.

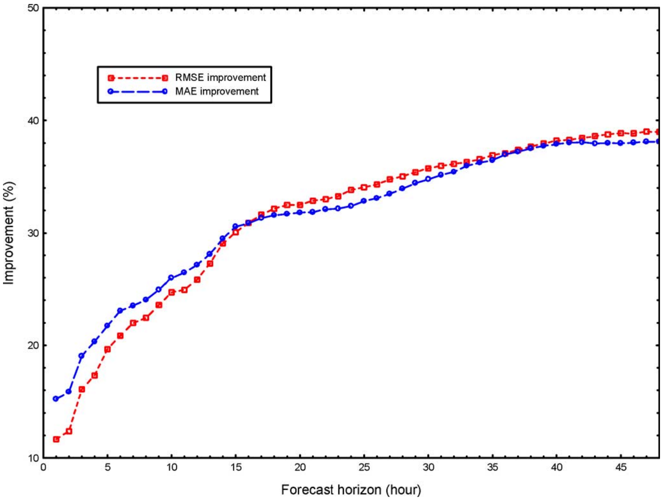

Fig. 9. Improvement over persistence for a horizon of 48 h obtained by the proposed model.

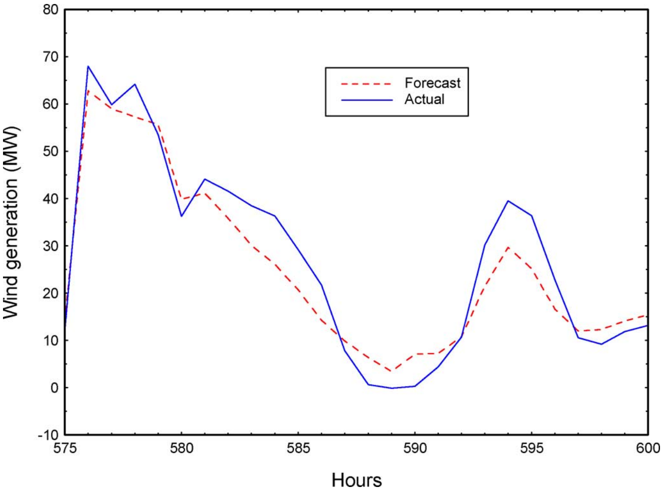

Fig. 10. Real and estimated wind power generation for a typical 24 h forecast from June 2005.

with persistent forecast, and considerable improvement up to more than 40% over the persistence is achieved.

Fig. 10 shows the forecasting and the actual wind power generation curves. To illustrate the variety of wind power more clearly, we select a typical 24-h forecast from the testing set for presentation. As shown in this figure, the proposed model accurately predicts the trend of the wind generation curve generally.

V. Conclusion

In this paper, an integrated machine learning forecasting model, based on BCD and SVR, has been developed to provide 48-h-ahead wind power generation forecasts for a wind farm. The proposed model has some notable advantages: first, it is practical and cost-efficient based on meteorological information that is easy to access; second, it can well handle the nonstationarity in the wind power and speed time series by using BCD to cluster the time series and multiple local powerful models—SVRs to fit the data in each subset; third, it has strong robustness and can be easily modified for different wind farms with minimum effort. Experiments and comparisons with the persistence method demonstrate the effectiveness and efficiency of the proposed model in learning and predicting wind generation.

Acknowledgment

The authors would like to thank Western Farmers Electric Cooperative for providing the data of the wind farm and WEATHERBANK, Inc., for providing the meteorological data.

References

[1] Global Wind 2006 Report, Global Wind Energy Council. Brussels, Belgium, 2006.

[2] J. O. G. Tande and L. Landberg, "A 10 sec. forecast of wind turbine output with neural networks," in *Proc. 4th Eur. Wind Energy Conf. (EWEC 1993)*, Lübeck-Travemünde, Germany, pp. 747–777.

[3] C. W. Potter and M. Negnevitsky, "Very short-term wind forecasting for tasmanian power generation," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 965–972, May 2006.

[4] G. Giebel, G. Kariniotakis, R. Brownsword. (2003). The state of the art in short-term prediction of wind power—A literature overview, *Position paper for the ANEMOS project* [Online]. Available http://www.anemos-project.eu

[5] G. N. Kariniotakis, G. S. Stavrakakis, and E. F. Nogaret, "Wind power forecasting using advanced neural networks models," *IEEE Trans. Energy Convers.*, vol. 11, no. 4, pp. 762–767, Dec. 1996.

[6] M. C. Alexiadis, P. S. Dokopoulos, and H. S. Sahsamanoglou, "Wind speed and power forecasting based on spatial correlation modes," *IEEE Trans. Energy Convers.*, vol. 14, no. 3, pp. 836–842, Sep. 1999.

[7] S. Li, D. C. Wunsch, E. A. O'hair, and M. G. Giesselmann, "Using neural networks to estimate wind turbine power generation," *IEEE Trans. Energy Convers.*, vol. 16, no. 3, pp. 276–282, Sep. 2001.

[8] 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, no. 2, pp. 352–361, Jun. 2004.

[9] T. G. Barbounis, J. B. Theocharis, M. C. Alexiadis, and P. S. Dokopoulos, "Long-term wind speed and power forecasting using local recurrent neural network models," *IEEE Trans. Energy Convers.*, vol. 21, no. 1, pp. 273–284, Mar. 2006.

[10] K. Methaprayoon, W. J. Lee, C. Yingvivatanapong, and J. R. Liao, "An integration of ANN wind power estimation into UC considering the forecasting uncertainty," in *Proc. IEEE Ind. Com. Power Syst. Tech. Conf.*, Saratoga Springs, NY, May 8–11, 2005, pp. 116–124.

[11] J. P. Eckmann, S. O. Kampshort, and D. Ruelle, "Recurrence plots of dynamical systems," *Europhys. Lett.*, vol. 4, pp. 973–977, 1987.

[12] J. Belaire-Franch and D. Contreras. (2002). "Recurrence plot in nonlinear time series analysis: Free software," *J. Stat. Softw.*, vol. 7, pp. 1–17 [Online]. Available http://home.netcom.com/∼eugenek/download.html

[13] M. Ramoni, P. Sebastiani, and P. Cohen, "Bayesian clustering by dynamics," *Mach. Learn.*, vol. 47, pp. 91–121, 2002.

[14] M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag, 1997.

[15] C. Cortes and V. Vapnik, "Support-vector network," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

[16] N. Cristianini and J. Shawe-Tylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[17] K.-R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in *Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 243–254.

[18] C.-C. Chang and C.-J. Lin. (2001). *LIBSVM: A Library for Support Vector Machines* [Online]. Available: http://www.csie.ntu.edu.tw/˜cjlin/libsvm

Shu Fan (M'08) received the B.S. M.S. and Ph.D. degrees from the Department of Electrical Engineering, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995, 2000, and 2004, respectively.

He was a postdoctoral researcher sponsored by the Japanese Government in Osaka Sangyo University from 2004 to 2006. From 2006 to 2007, he was a Visiting Scholar at Energy Systems Research Center, University of Texas, Arlington. He is currently a Research Fellow at Monash University, Clayton, Vic., Australia. His current research interests include energy system forecasting, power system control, and high-power electronics.

James R. Liao (M'89) received the M.S. degree from the University of Missouri, Rolla, MO, in 1980, and the Ph.D. degree from the University of Oklahoma, Norman, OK, in 1992, all in electrical engineering.

Since 1980, he has been with the Western Farmers Electric Cooperative, Anadarko, OK, where he was a Transmission/Generation Systems Analyst from 2006 to 2007, an EMS System Software Engineer from 1985 to 1999, and the Principal Operations Engineer since 1999.

Dr. Liao is a Registered Professional Engineer in the State of Oklahoma.

Ryuichi Yokoyama (M'82–SM'06) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Waseda University, Tokyo, Japan, in 1968, 1970, and 1973, respectively.

He was a Professor in the Faculty of Engineering, Tokyo Metropolitan University, Hachioji, Tokyo. He is currently a Professor at Waseda University. His current research interests include planning, operation, control, and simulation of large-scale power systems, and also analysis of deregulated electricity markets.

Prof. Yokoyama is a member of the Society of Instrument and Control Engineers (SICE), Japan, and the International Council on Large Electric System (CIGRE), Paris, France.

Luonan Chen (M'94–SM'98) received the B.E. degree from Huazhong University of Science and Technology (HUST), Wuhan, China, in 1984, and the M.E. and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1988 and 1991, respectively.

Since 1997, he has been a faculty of Osaka Sangyo University, Osaka, Japan, where he is currently a Professor in the Department of Electrical Engineering and Electronics. His current research interests include nonlinear dynamics and optimization in power systems.

Wei-Jen Lee (S'85–M'85–SM'97–F'07) received the B.S. and M.S. degrees from National Taiwan University, Taipei, Taiwan, R.O.C., and the Ph.D. degree from the University of Texas, Arlington, in 1978, 1980, and 1985, respectively, all in electrical engineering.

In 1985, he joined the University of Texas, where he is currently a Professor in the Department of Electrical Engineering and the Director of the Energy Systems Research Center. His current research interests include research on power flow, transient, and dynamic stability, voltage stability, short circuits, relay coordination, power quality analysis, and deregulation for utility companies.

Dr. Lee is a Registered Professional Engineer in the State of Texas.