

## Research Article

# Day-Ahead Wind Speed Forecasting Using Relevance Vector Machine

Guoqiang Sun,<sup>1</sup> Yue Chen,<sup>1</sup> Zhinong Wei,<sup>1</sup> Xiaolu Li,<sup>2</sup> and Kwok W. Cheung<sup>3</sup>

<sup>1</sup> Research Center for Renewable Energy Generation Engineering, Ministry of Education, Hohai University, Nanjing 210098, China

<sup>2</sup> ALSTOM GRID Technology Center Co., Ltd., Shanghai 201114, China

<sup>3</sup> ALSTOM Grid Inc., Redmond, WA 98052, USA

Correspondence should be addressed to Guoqiang Sun; [hhusunguoqiang@163.com](mailto:hhusunguoqiang@163.com)

Received 31 December 2013; Revised 5 May 2014; Accepted 22 May 2014; Published 12 June 2014

Academic Editor: Hongjie Jia

Copyright © 2014 Guoqiang Sun et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the development of wind power technology, the security of the power system, power quality, and stable operation will meet new challenges. So, in this paper, we propose a recently developed machine learning technique, relevance vector machine (RVM), for day-ahead wind speed forecasting. We combine Gaussian kernel function and polynomial kernel function to get mixed kernel for RVM. Then, RVM is compared with back propagation neural network (BP) and support vector machine (SVM) for wind speed forecasting in four seasons in precision and velocity; the forecast results demonstrate that the proposed method is reasonable and effective.

## 1. Introduction

Over the past decade, people in many countries worldwide have paid significant attention to wind power generation because of it being pollution-free, clean, and renewable. At the end of 2012, worldwide installed capacity of wind power reached 282.2 GW, increased almost 20% compared to the previous year 2011 which of 240 GW. By the end of 2020, total global installed capacity will reach 1150 GW, and wind power will be over 2800 TWh, accounting for about 12% of global electricity demand; by the end of 2030, installed capacity will exceed 2500 GW, and wind power generating capacity will reach 6600 TWh, accounting for about 23% of global electricity demand [1]. The introduction of such a large-scale wind power has attracted many domestic and foreign scholars for further wind power technology. The wind forecast as a basic link of the wind power research is one of the effective ways to solve the problem and has an important role in the safe and economic operation of the power grid, so a growing number of researchers pay attention to it recently.

We can cluster the wind forecasting techniques into two main groups; the first group are physical methods, taking physical considerations into account, such as temperature and local terrain. In [2, 3], numerical weather prediction

(NWP) model could be used directly for wind speed and wind energy predictions.

Another group are statistical methods. Conventional ones are identical to the direct random time-series model, such as autoregressive model (AR), moving average model (MA), autoregressive moving average model (ARMA), and autoregressive integrated moving average model (ARIMA). Kamal and Jafri [4] established an ARMA model and found for long-term or short-term predictions, the values of variances and wind speed with a confidence interval of 95% were acceptable. A fractional-ARIMA (f-ARIMA) model was used by Kavasseri and Seetharaman [5] for day-ahead and two-day-ahead wind speed forecasting. Results showed that forecast accuracy was significantly improved with f-ARIMA model compared to the previous method.

Apart from the mentioned forecasting techniques, machine learning algorithms such as artificial neural network (ANN), Bayesian network (BN), and support vector machine (SVM) are usually adopted for time series-based wind prediction. Bilgili et al. [6] investigated the use of a model based on the ANN method and spatial correlation for monthly wind speed prediction without any topographic details or other meteorological data. The prediction results showed that the maximum MAE was 14.13%, while

the minimum was 4.49% in the developed model. Welch [7] compared three types of neural networks (multilayer perceptron neural network, simultaneous recurrent neural network, and Elman recurrent neural network) for short-term prediction of wind speed, and all training data used particle swarm optimization (PSO). Mohandes et al. [8] established a SVM model to predict wind speed, compared with a multilayer perceptron ANN model. The results showed that SVM model gave lower root mean square error than the MLP ANN model. Larson and Westrick [9] used a support vector classifier to estimate the forecasting error, obtaining lower mean square error and mean absolute percentage error than traditional SVM.

However, the SVM had a number of significant and practical limitations. For example, we could not get probabilistic predictions and the kernel function must satisfy Mercer's condition. In order to overcome these, Tipping [10] and Tipping and Faul [11] proposed an advanced function estimation technique, relevance vector machine (RVM). RVM utilizes a more flexible and sparser function without extra regularization parameters; it is an inherent machine learning technique [12].

This paper establishes RVM model to forecast day-ahead wind speed in JiangSu, compared with BP and SVM models. It can show that the proposed method is more effective and robust and gets rid of the overfitting problem of traditional nonparametric regression models. The rest of the paper is organized as follows. A brief review of the theory of RVM learning for classification is provided in Section 2. Then, in Section 3, several forecasting examples are given to contrast RVM model with other different models. Finally Section 4 will give the conclusions.

## 2. Classical Relevance Vector Machine

Relevance vector machine (RVM), based on the overall Bayesian framework, is a sparse probability model and now is one of the hot research fields [13]. The principle is as follows.

Given a training dataset  $\{x_i, t_i\}_{i=1}^N$  ( $i = 1, 2, \dots, N$ ), where  $\{x_i\}_{i=1}^N$  are the  $N$ -dimensional input vectors and  $\{t_i\}_{i=1}^N$  are output vectors. Here we assume that the targets are independent and contaminated with zero mean random error (noise) vector  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)^T$ , where  $\varepsilon_i \sim \mathbf{N}(0, \sigma^2)$  ( $i = 1, 2, \dots, N$ ) and  $\mathbf{N}(0, \sigma^2)$  is the normal distribution with mean 0 and variance  $\sigma^2$ . The output can be defined as

$$t_i = y(x_i, \boldsymbol{\omega}) + \varepsilon_i, \quad (1)$$

where the weights  $\boldsymbol{\omega} = (\omega_0, \omega_1, \dots, \omega_N)^T$  and  $y(\mathbf{x}, \boldsymbol{\omega})$ , as a linearly-weighted sum of  $\mathbf{N}$ , is generally fixed and nonlinear.

Consider

$$y(\mathbf{x}, \boldsymbol{\omega}) = \sum_{i=1}^N \omega_i K(x, x_n) + \omega_0 = \boldsymbol{\Phi} \boldsymbol{\omega}, \quad (2)$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_N) \\ 1 & K(x_2, x_1) & K(x_2, x_2) & \cdots & K(x_2, x_N) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & K(x_N, x_1) & K(x_N, x_2) & \cdots & K(x_N, x_N) \end{bmatrix}, \quad (3)$$

where  $K(x, x_n)$  is a kernel function.

Therefore, the probability formula for relevance vector machine model is

$$p(t_i | \mathbf{x}) \sim \mathbf{N}(y(x_i), \sigma^2), \quad (4)$$

where  $\mathbf{N}(y(x_i), \sigma^2)$  is the normal distribution with mean  $y(x_i)$  and variance  $\sigma^2$ .

Because we assume the targets are independent, thus the likelihood of the complete dataset can be defined as

$$p(\mathbf{t} | \mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{w}\|^2\right\}. \quad (5)$$

Because there are so many parameters in the model as training examples, maximum likelihood estimation of  $\mathbf{w}$  and  $\sigma^2$  would easily lead to overfitting. To avoid this, the sparse Bayesian principles are utilized to give  $\mathbf{w}$  zero-mean Gaussian prior distribution:

$$p(\mathbf{w} | \boldsymbol{\alpha}) = \prod_{i=0}^N \mathbf{N}(\omega_i | \mathbf{0}, \alpha_i^{-1}), \quad (6)$$

where  $\boldsymbol{\alpha}$  are the  $N + 1$  dimensional vectors and each weight separately corresponds to a hyperparameter which controls how far from zero each weight is allowed to deviate [13]. Having defined the prior, using Bayes' rule, noninformative prior distribution can be given by

$$p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \sigma^2 | \mathbf{t}) = \frac{p(\mathbf{t} | \boldsymbol{\omega}, \boldsymbol{\alpha}, \sigma^2) p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \sigma^2)}{\int p(\mathbf{t} | \boldsymbol{\omega}, \boldsymbol{\alpha}, \sigma^2) p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \sigma^2) d\boldsymbol{\omega} d\boldsymbol{\alpha} d\sigma^2}. \quad (7)$$

However, we cannot easily gain the full analytical solution to (7); thus, we decompose the posterior according to  $p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \sigma^2 | \mathbf{t}) = p(\boldsymbol{\omega} | \mathbf{t}, \boldsymbol{\alpha}, \sigma^2) p(\boldsymbol{\alpha}, \sigma^2 | \mathbf{t})$  and get the solution to this integral (7). Then, the posterior distribution over the weights can be written by

$$p(\boldsymbol{\omega} | \mathbf{t}, \boldsymbol{\alpha}, \sigma^2) = \frac{p(\mathbf{t} | \boldsymbol{\omega}, \sigma^2) p(\boldsymbol{\omega} | \boldsymbol{\alpha})}{p(\mathbf{t} | \boldsymbol{\alpha}, \sigma^2)}. \quad (8)$$

After defining the prior distribution and the likelihood distribution, according to Bayes' theorem, we can obtain the posterior distribution of all unknown parameters:

$$p(\boldsymbol{\omega} | \mathbf{t}, \boldsymbol{\alpha}, \sigma^2) = \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (9)$$

Then the mean and the covariance of  $\mathbf{N}(\boldsymbol{\mu}, \Sigma)$  are expressed as

$$\Sigma = (\sigma^{-2} \Phi^T \Phi + A)^{-1}, \quad (10)$$

$$\boldsymbol{\mu} = \sigma^{-2} \Sigma \Phi^T \mathbf{t}, \quad (11)$$

where  $A = \text{diag}(\alpha_0, \alpha_1, \dots, \alpha_N)$ .

In order to calculate  $\boldsymbol{\alpha}$  and  $\sigma^2$ , using the maximum likelihood calculation, we need to only maximize the term  $p(\mathbf{t} | \boldsymbol{\alpha}, \sigma^2)$ :

$$\begin{aligned} p(\mathbf{t} | \boldsymbol{\alpha}, \sigma^2) &= \int p(\mathbf{t} | \boldsymbol{\omega}, \sigma^2) p(\boldsymbol{\omega} | \boldsymbol{\alpha}) d\boldsymbol{\omega} \\ &= \left( \frac{(2\pi)^{-N/2}}{\sqrt{|\Omega|}} \right) \exp\left(-\frac{1}{2} \mathbf{t}^T \Omega^{-1} \mathbf{t}\right), \end{aligned} \quad (12)$$

where  $\Omega = \sigma^2 + \Phi A^{-1} \Phi^T$ .

In related Bayesian models, this quality is referred to as “the marginal likelihood” [14, 15] or the “evidence for hyperparameter” [16]; its maximization is known as the “type II maximum likelihood method” or “evidence procedure.”

Because we cannot obtain values of  $\boldsymbol{\alpha}$  and  $\sigma^2$  which maximize (12) in closed form, so here an iterative reestimation method is needed. We partially differentiated (12) and made  $\boldsymbol{\alpha}$  equal to 0; then, the following approach gives [17]

$$\frac{p(\mathbf{t} | \boldsymbol{\alpha}, \sigma^2)}{d\boldsymbol{\alpha}} = 0 \longrightarrow \alpha_i^{\text{new}} = \frac{\gamma_i}{\mu_i^2}, \quad (13)$$

where  $\mu_i$  is the posterior mean weight at time of  $i$  from (11), and the quantity  $\gamma_i$  can be defined as

$$\gamma_i \equiv 1 - \alpha_i \Sigma_{ii}, \quad (14)$$

where the  $\Sigma_{ii}$  is the diagonal element at time of  $i$  of the posterior weight covariance (10).

When  $\alpha_i$  is large,  $\omega_i$  is highly constrained by the prior, and then  $\Sigma_{ii} \approx \alpha_i^{-1}$  and  $\gamma_i \approx 0$ ; conversely, when  $\alpha_i$  is small, we can know that  $\gamma_i \approx 1$ . So  $\gamma_i$  can be interpreted as the intermediate variables measuring the best parameter  $\alpha_i$ .

Meanwhile, for the noise variance  $\sigma^2$ , differentiation of (12), equating to 0, then the following approach gives [16]

$$\frac{p(\mathbf{t} | \boldsymbol{\alpha}, \sigma^2)}{d\sigma^2} = 0 \longrightarrow (\sigma^2)^{\text{new}} = \frac{\|\mathbf{t} - \Phi \boldsymbol{\mu}\|^2}{N - \sum_{i=0}^N \gamma_i}, \quad (15)$$

where  $N$  is the number of data examples.

In practice, because many of  $\alpha_i$  are found to tend to infinity during reestimation, the posterior distribution from formula (9) of the corresponding weights  $w_i$  will highly peak at zero. In this optimization process, those  $w_i$  become zero, and the vectors from the training set that associate with the remaining nonzero weights  $w_i$  are called relevance vectors.

At convergence of the hyperparameter estimation procedure, we use the posterior distribution over the weights for predictions, conditioned on the maximizing values  $\boldsymbol{\alpha}_{MP}$  and

$\sigma_{MP}^2$ , which are optimum values used for prediction through iteration. Then the predictive distribution can be written as

$$\begin{aligned} p(\mathbf{t}_* | \mathbf{t}, \boldsymbol{\alpha}_{MP}, \sigma_{MP}^2) \\ = \int p(\mathbf{t}_* | \boldsymbol{\omega}, \sigma_{MP}^2) p(\boldsymbol{\omega} | \mathbf{t}, \boldsymbol{\alpha}_{MP}, \sigma_{MP}^2) d\boldsymbol{\omega}. \end{aligned} \quad (16)$$

Because both terms in the integrand obey Gaussian distribution, we can easily compute the predictive distribution:

$$p(\mathbf{t}_* | \mathbf{t}, \boldsymbol{\alpha}_{MP}, \sigma_{MP}^2) = \mathbf{N}(\mathbf{t}_* | \mathbf{y}_*, \sigma_*^2), \quad (17)$$

where the new input values are  $\mathbf{x}_*$  and

$$\mathbf{y}_* = \boldsymbol{\mu}^T \Phi(\mathbf{x}_*), \quad (18)$$

$$\sigma_*^2 = \sigma_{MP}^2 + \Phi(\mathbf{x}_*)^T \Sigma \Phi(\mathbf{x}_*). \quad (19)$$

Eventually, we get the regression model of RVM in function (18) and the variance function in function (19).

In summary, the forecast process can be summarized as in the following steps:

- (1) initialize variances  $\sigma^2$  and hyperparameters  $\alpha_i$ ;
- (2) compute posteriori statistics of weights  $\boldsymbol{\mu}$  and  $\Sigma$ ;
- (3) compute  $\gamma_i$  and update the  $\alpha_i$  and  $\sigma^2$ ;
- (4) if it is not convergent, then go back step (2); otherwise, go to step (5);
- (5) if  $\alpha_i$  tend to infinity, the weights  $\omega_i$  can be deleted;
- (6) get the predictive mean intuitively from  $y(\mathbf{x}_*, \boldsymbol{\mu})$ .

### 3. Illustrative Examples

**3.1. Data and Pretreatment.** In this study, wind speed values throughout 2008 on a wind farm in Jiangsu are taken as the training samples, and all the data have an interval period of 15 min. To evaluate the performance of proposed model, we establish RVM model for 96 points wind speed prediction, compared with SVM and BP in terms of forecast accuracy, model running time, and model complexity. Based on historical data in each quarter, we establish forecasting models to predict day-ahead wind speed on March 25, June 26, September 29, and December 28.

Because the input vectors contain different kinds of physical quantities, in order to ensure the variables are comparable, but also solve problems such as the increasing of training time, first of all, we take normalization process on all input data. The normalized respective variables will be in  $[0, 1]$

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \quad (20)$$

where  $x_{\max}$  and  $x_{\min}$  are, respectively, the maximum value and the minimum value for each variable.

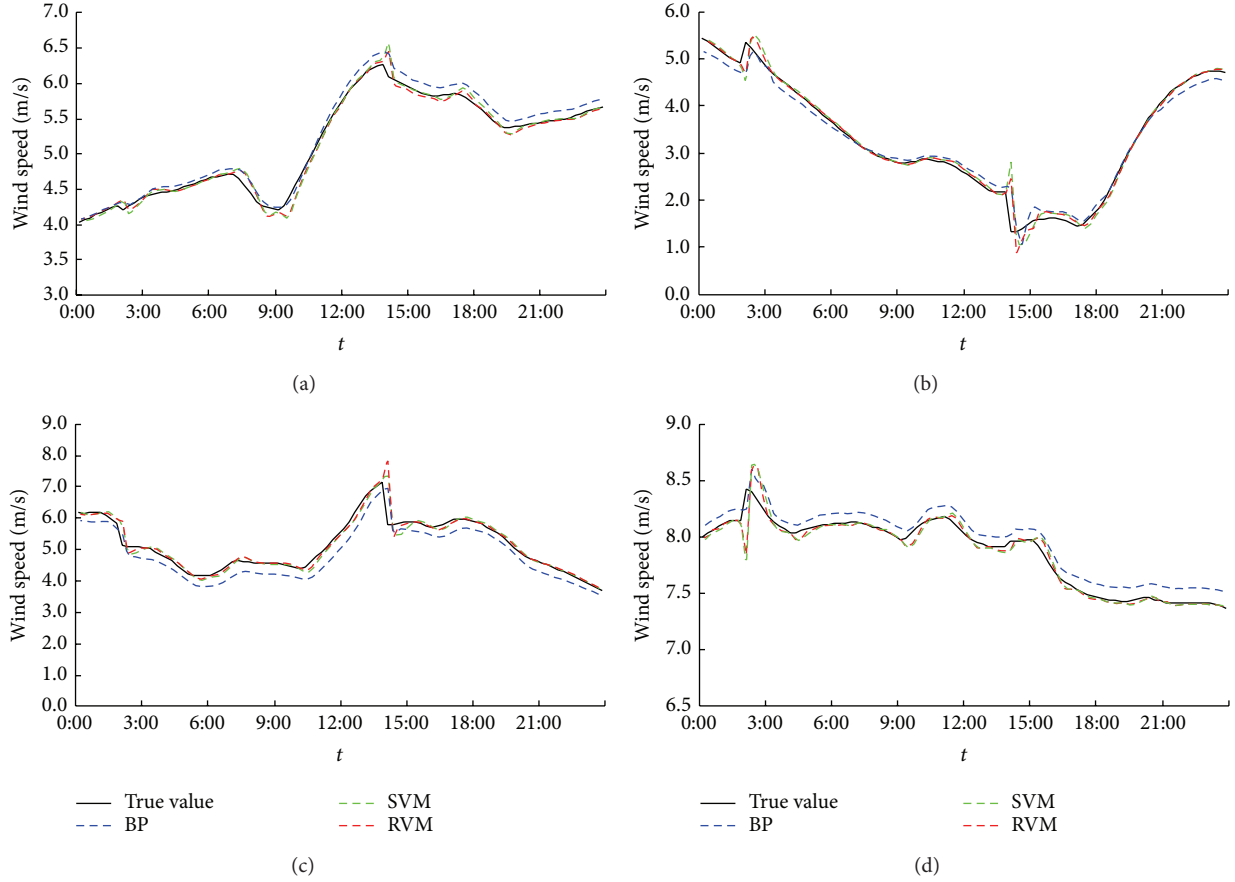


FIGURE 1: Forecasting results with different models on March 25 (a), June 26 (b), September 29 (c), and December 28 (d).

For evaluating the forecasting performance, mean relative error (MRE) and root mean square error (RMSE) are used; they are defined in functions

$$e_{\text{MRE}} = \frac{1}{n} \sum_{i=1}^n \frac{|Y_i - \hat{Y}_i|}{|Y_i|}, \quad (21)$$

$$e_{\text{RMSE}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2},$$

where  $Y_i$  and  $\hat{Y}_i$  indicate actual and forecasting values of wind speed at time of  $i$ .

**3.2. Select Kernel Function of RVM.** Because the kernel function of RVM does not need to satisfy Mercer's condition, the selection of the kernel function has a certain degree of freedom. The basic idea of the hybrid kernel function [17] is that several kernel functions can be combined together for different nature, and the new properties can be integrated and reflected.

TABLE 1: Comparisons of forecast accuracy for three models.

Month	BP		SVM		RVM	
	RMSE	MRE	RMSE	MRE	RMSE	MRE
March	0.1123	0.0180	0.0925	0.0115	0.0782	0.0108
June	0.1856	0.0520	0.2033	0.0428	0.1607	0.0347
September	0.3521	0.0641	0.2297	0.0230	0.2451	0.0185
December	0.1106	0.0136	0.0930	0.0057	0.0790	0.0046
Avg.	<b>0.1902</b>	<b>0.0369</b>	<b>0.1546</b>	<b>0.0208</b>	<b>0.1408</b>	<b>0.0172</b>

In this paper, we combine Gaussian kernel function and polynomial kernel function to get mixed kernel of global and local nature:

$$K(x, x_i) = \lambda G(x, x_i) + (1 - \lambda) P(x, x_i),$$

$$G(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{\sigma^2}\right), \quad (22)$$

$$P(x, x_i) = (x \cdot x_i + 1)^2,$$

where  $G(x, x_i)$  is RBF kernel function;  $P(x, x_i)$  is binomial kernel function;  $\lambda$  is the weight of the kernel function; and  $\sigma$

TABLE 2: Comparisons of test time and vector number for three models.

Month	BP		SVM		RVM	
	Test time (s)	Number of vectors or neurons involved	Test time (s)	Number of vectors or neurons involved	Test time (s)	Number of vectors or neurons involved
March	677.56	300	639.24	216	302.48	33
June	1237.43	380	857.25	306	288.87	40
September	1002.86	340	641.91	320	293.85	35
December	835.03	320	506.29	202	288.66	31
Avg.	<b>938.21</b>	<b>335</b>	<b>661.17</b>	<b>261</b>	<b>293.47</b>	<b>34.75</b>

is the kernel width. There, the grid search method is employed to get the optimal values of  $\lambda$  and  $\sigma$ .

**3.3. Different Forecasting Models.** To evaluate the performance of proposed model, RVM, BP, and SVM are established for day-ahead wind speed predictions on March 25, June 26, September 29, and December 28. Table 1 shows the comparisons of forecast accuracy for three models in each quarter, and Table 2 shows the comparisons of test time and number of vectors or neurons for three models. The specific forecasting results are presented in Figure 1.

It is found from the comparison results that the forecast accuracy of the proposed method is higher than that used by the other models. The average RMSE of RVM model is only 0.1408 m/s, lower than those of BP and SVM, which are 0.1902 m/s and 0.1546 m/s. The average MRE of RVM is 1.72%, while those of BP and SVM are 3.69% and 2.08%. In different seasons, forecast accuracy is different. In spring and winter, wind speed changes are relatively small, so higher precision can be got; in summer and autumn, wind speed at the coast has large fluctuations, so the accuracy will decrease.

Table 2 shows clearly that RVM model can reduce model complexity; the average number of vectors or neurons involved is only 34.75, and comparatively those of BP and SVM are 335 and 261. Although RVM typically utilizes dramatically fewer kernel functions, its generalization performance is comparable to SVM. From Table 2, we know that RVM has a higher sparse network and can quickly converge; its test time is 293.47 s, shorter than those of BP (938.21 s) and SVM (661.17 s) models.

## 4. Conclusions

In this paper, RVM is proposed for day-ahead wind speed forecasting. Firstly, we combine Gaussian kernel function and polynomial kernel function to get mixed kernel for RVM model. Then, we compare RVM with BP and SVM for wind speed forecasting in four seasons for precision and velocity. Finally, the simulation results show that the proposed method is more effective and robust and has better performance in terms of forecast accuracy, model running time, and model complexity than that used by BP neural network and SVM model. So the theoretical feasibility of RVM for the wind speed prediction has the some meaning.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 51277052, 51107032, and 61104045.

## References

- [1] Global Wind Energy Council and Greenpeace, *Global Wind Power Development Outlook in 2012*, Global Wind Energy Council and Greenpeace, Beijing, China, 2012.
- [2] M. S. Roulston, D. T. Kaplan, J. Hardenberg, and L. A. Smith, "Using medium-range weather forecasts to improve the value of wind energy production," *Renewable Energy*, vol. 28, no. 4, pp. 585–602, 2003.
- [3] L. Lazić, G. Pejanović, and M. Živković, "Wind forecasts for wind power generation using the Eta model," *Renewable Energy*, vol. 35, no. 6, pp. 1236–1243, 2010.
- [4] L. Kamal and Y. Z. Jafri, "Time series models to simulate and forecast hourly averaged wind speed in Quetta, Pakistan," *Solar Energy*, vol. 61, no. 1, pp. 23–32, 1997.
- [5] R. G. Kavasseri and K. Seetharaman, "Day-ahead wind speed forecasting using f-ARIMA models," *Renewable Energy*, vol. 34, no. 5, pp. 1388–1393, 2009.
- [6] M. Bilgili, B. Sahin, and A. Yasar, "Application of artificial neural networks for the wind speed prediction of target station using reference stations data," *Renewable Energy*, vol. 32, no. 14, pp. 2350–2360, 2007.
- [7] R. L. Welch, S. M. Ruffing, and G. K. Venayagamoorthy, "Comparison of feedforward and feedback neural network architectures for short term wind speed prediction," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '09)*, pp. 3335–3340, Atlanta, Ga, USA, June 2009.
- [8] M. A. Mohandes, T. O. Halawani, S. Rehman, and A. A. Hussain, "Support vector machines for wind speed prediction," *Renewable Energy*, vol. 29, no. 6, pp. 939–947, 2004.
- [9] K. A. Larson and K. Westrick, "Short-term wind forecasting using off-site observations," *Wind Energy*, vol. 9, no. 1-2, pp. 55–62, 2006.
- [10] M. E. Tipping, "The relevance vector machine," in *Advances in Neural Information Processing Systems*, vol. 12, pp. 652–658, 2000.

- [11] M. E. Tipping and A. C. Faul, "Fast marginal likelihood maximization for sparse Bayesian models," in *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, vol. 1, Key West, Fla, USA, January 2003.
- [12] P. K. Wong, Q. Xu, C. M. Vong, and H. C. Wong, "Rate-dependent hysteresis modeling and control of a piezostage using online support vector machine and relevance vector machine," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 4, pp. 1988–2001, 2012.
- [13] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 211–244, 2001.
- [14] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer Series in Statistics, Springer, New York, NY, USA, 1985.
- [15] G. Wahba, "A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem," *The Annals of Statistics*, vol. 13, no. 4, pp. 1378–1402, 1985.
- [16] D. J. MacKay, *Bayesian methods for adaptive models [Ph.D. thesis]*, California Institute of Technology, 1992.
- [17] Q. Duan, J. G. Zhao, and Y. Ma, "Relevance vector machine based on particle swarm optimization of compounding kernels in electricity load forecasting," *Electric Machines and Control*, vol. 14, no. 6, pp. 33–38, 2010.