



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/135733/>

Version: Accepted Version

Article:

Yan, J, Li, K, Bai, E et al. (2016) Time series wind power forecasting based on variant Gaussian Process and TLBO. *Neurocomputing*, 189. pp. 135-144. ISSN: 0925-2312

<https://doi.org/10.1016/j.neucom.2015.12.081>

© 2016 Elsevier B.V. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Time series wind power forecasting based on variant Gaussian Process and TLBO

Juan Yan^a, Kang Li^{a,*}, Erwei Bai^{a,b}, Zhile Yang^a, Aoife Foley^c

^a*School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT7 1NN*

^b*Department of Electrical and Computer Engineering, University of Iowa, Iowa, IA 52242, USA*

^c*School of Mechanical and Aerospace Engineering, Queen's University Belfast, BT7 1NN*

Abstract

Due to the variability and stochastic nature of wind power, accurate wind power forecasting plays an important role in developing reliable and economic power system operation and control strategies. As wind variability is stochastic, Gaussian Process regression has recently been introduced to capture the randomness of wind energy. However, the disadvantages of Gaussian Process regression include its computation complexity and incapability to adapt to time varying time-series systems. A variant Gaussian Process for time series forecasting is introduced in this study to address these issues. This new method is shown to be capable of reducing computational complexity and increasing prediction accuracy. It is further proved that the forecasting result converges as the number of available data approaches infinite. Further, a teaching learning based optimization (TLBO) method is used to train the model and to accelerate the learning rate. The proposed modelling and optimization method is applied to forecast both the wind power generation in Ireland and that from a single wind farm to demonstrate the effectiveness of the proposed method.

Keywords: Gaussian Process, Model Consistency, TLBO, Wind Power Forecasting

*Corresponding author

Email address: k.li@qub.ac.uk (Kang Li)

1. Introduction

As power systems in many countries and regions are penetrated with increasing wind power, it is imperative to forecast wind power generation accurately in advance for reliable and effective power system operation and control. Currently, wind energy time series forecasting has shown to be an effective technique for short term forecasting [1]. Unlike the numerical weather prediction (NWP) methods, which employ weather information such as temperature, wind speed, wind direction etc, time series models employ historical measurement data solely, to make short term predictions from several minutes to several hours ahead, which could be very useful for short term load balancing and energy storage decisions [2]. Although such forecasting horizon is relatively short in comparison with NWP, time series methods have demonstrated their great advantage in saving computation resources. Existing time series forecasting techniques include traditional methods such as ARMA [3], Persistence, Neural Network [4][5][6], Neural-fuzzy [7] etc. Besides, methods such as Kalman filters [8] and Gaussian Process (GP) [9][10][11] have also been recently introduced.

Although the GP approach was first used in the statistics community in 1964 [12] and applied to curve fitting in 1978 [13], this stochastic process did not attract much attention until a comparison between GP and other well known methods was carried out by Ramsmussen in 1996 [14]. Since then, the implementing and application of GP have been further researched and extended. Initially the GP learning process was studied and simplified [15], then GP was applied to system regression [16, 17], and classification [18, 19] etc. GP is a global non-parametric method that assumes that all the variables follow one joint Gaussian distribution and all the available data are employed in the prediction procedure. It is a special case of Bayesian inference, where all the priori are assigned to be Gaussian. Moreover, GP is viewed as similar to kernel estimators, because a covariance function is used to describe the correspondence between two outputs. The main difference between GP and kernel estimators lies in that the sum of the weights does not have to be unity. Alternatively,

GP can be viewed as a form of basis function approach, differing from those normal ones due to the flexible coefficients [20]. One significant advantage of GP is that besides giving the most probable estimation, GP describes the distribution of the new prediction which can be quite beneficial in developing model based control strategies in industry. Secondly, the global property of GP guarantees robust estimation even when the number of available data is limited or imbalanced. Moreover, its covariance function contains less hyperparameters in comparison with other advanced machine learning methods such as Neural Network, and Fuzzy logic etc., and thus avoids the curse of dimensionality as the dimension of input increases.

Due to these advantages, GP regression began to be applied in a variety of fields, from multi sensor networks [21, 22] to image processing [23, 24, 25], from semiconductor industrial process [26] to medical health [27] and biological observation [28]. However, drawbacks still exist. First, in GP all available data is assumed to follow one joint Gaussian distribution and employed further to make new predictions. Such mechanisms generate expensive computational demand caused by the matrix inversion in GP modelling. Especially less relevant data being used a good prediction while causing unnecessary computational burden. Secondly, its ability to reflect the local property of a system remains to be an issue. To tackle these problems, sparse approximation techniques for full GP [29, 30] and methods of local GP mixtures [31] have been proposed recently. Further, GP methods have been applied in wind power forecasting by Yan et al [32].

In this paper, a variant GP for time series system is developed and the model consistency is proved using a test theory. In comparison with other existing variant GP models, the method proposed in this paper emphasizes on the underlying temporally local property of acquired data from the time-varying wind power systems and shows great accuracy in the real application to both the all-island wind generation and a small farm output. Moreover, the optimization techniques are investigated and a new optimization technique, namely the teaching-learning based optimization (TLBO) is used, to overcome the limita-

tions of some conventional optimization techniques such as linear programming and quadratic programming. TLBO is a new member of meta-heuristic optimization family proposed in 2011 [33] and it has been adopted to solve a number of real-world problems [34, 35, 36, 37] due to its fast convergence speed and excellent exploitation ability. Therefore, TLBO is used to optimize the proposed GP variant model and compared to other heuristic methods.

This paper is organized as follows. First, the standard GP for time series wind power forecasting is presented in Section II. In Section III, the variant efficient GP is presented in detail, together with the computational analysis and the model convergence proof. Following the model description, the learning and teaching procedure of TLBO is introduced in Section IV. In Section V, the wind power generation for the whole island of Ireland and from a small farm on it are used as case studies to confirm the effectiveness of the proposed method. Finally, Section VI concludes the paper.

2. Gaussian Process for time series forecasting

2.1. Standard Gaussian Process

For a multiple-input-single-output (MISO) nonlinear system, let (X, Y) denote a set of input-output data \mathcal{D} , and suppose the k^{th} ($k \in [1, N]$) sample $(\mathbf{x}(k), y(k))$ satisfies equation (1), where v is an i.i.d random sequence of white noise with zero mean and finite variance σ_v^2 , which in the case of wind power forecasting refers to wind power measurement noise. Here $\mathbf{x} \in R^D$, which is the input vector.

$$y(k) = f(\mathbf{x}(k)) + v(k) \tag{1}$$

GP is a stochastic process where an indexed collection of random variables follow joint Gaussian distribution [38]. Generally, the mean function could be assumed to be zero if the data are properly scaled and de-trended [39] as shown

in equation (2).

$$P(Y|C_Y, X) = \mathcal{N}(0, C_Y) \quad (2)$$

For a given new output $y_0 = f(\mathbf{x}_0)$, if it follows joint Gaussian distribution with the available data Y in (2), then the joint distribution could be written as
 90 (3) in a partitioned form where A, B, C is shown in (4). Here, $cov(a, b)$ denotes the covariance between two variables a and b , and its value is decided by the so called covariance function.

$$\begin{bmatrix} y_0 \\ Y \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} A & B \\ B^\top & C_Y \end{bmatrix}\right) \quad (3)$$

$$A = cov(y_0, y_0)$$

$$B(i) = cov(y_0, y(i)) \quad y(i) \in Y \text{ and } i \in (1, N) \quad (4)$$

$$C_Y(i, j) = cov(y(i), y(j)) \quad y(i), y(j) \in Y \quad i, j \in (1, N)$$

There exist many forms of covariance functions [14]. The square exponential function shown in (5) is one of the most popular ones due to its infinite
 95 differentiability.

$$\begin{aligned} cov(y(i), y(j)) &= \Phi(\mathbf{x}(i), \mathbf{x}(j)) \\ &= s \cdot \exp\left[-\frac{1}{2} \sum_{d=1}^D \omega_d (x_d(i) - x_d(j))^2\right] + v \cdot \delta_{ij} \end{aligned} \quad (5)$$

Here, D refers to the dimension of model input \mathbf{x} and δ_{ij} refers to the Kronecker delta representing the observation noise for each sample, and Φ represents the covariance function. The hyperparameters involved could be denoted as $\theta = [s, v, \omega_1, \dots, \omega_D]$.

$$\begin{aligned} \ln P(Y|X, \theta^*) &= \ln \left[\frac{1}{(2\pi)^{\frac{N}{2}} |C_Y|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} Y^\top C_Y^{-1} Y\right) \right] \\ &= -\frac{1}{2} Y^\top C_Y^{-1} Y - \frac{1}{2} \ln |C_Y| - \frac{N}{2} \ln 2\pi \end{aligned} \quad (6)$$

100 Standard GP employs the gradient based methods to optimize the marginal likelihood function, due to its fast convergence rate and satisfactory accuracy. The gradient form of the log marginal density is shown in (7), where θ_j^* represents the j th element of the hyperparameter vector.

$$\frac{\partial P(Y|X, \theta^*)}{\partial \theta_j^*} = \frac{1}{2} Y^\top C_Y^{-1} \frac{\partial C_Y}{\partial \theta_j^*} C_Y^{-1} Y - \frac{1}{2} \text{Tr} \left[C_Y^{-1} \frac{\partial C_Y}{\partial \theta_j^*} \right] \quad (7)$$

2.2. Time series wind power forecasting

105 Time series prediction is an effective way for short term wind power forecasting. It employs only the historical measurement data and neglects the potential exogenous inputs, thus a time series system could be expressed in (8) where L represents the time lag.

$$y(t) = f(y(t-1), y(t-2), \dots, y(t-L)) + v(t) \quad (8)$$

Denote $\mathbf{x}(t) = [y(t-1), y(t-2), \dots, y(t-L)]^\top$, which represents the state vector at time t . In this case, L is equivalent to D in the previous section. Given a sequence of data Y as training data, for time instant t , the output could be predicted with (9) where $B(t)$ describes the covariance between $y(t)$ and Y , and C_Y denotes the self covariance of data sequence Y .

$$\hat{y}(t) = B(t) C_Y^{-1} Y \quad (9)$$

$$B(t) = (\Phi(\mathbf{x}(t), \mathbf{x}(1)), \Phi(\mathbf{x}(t), \mathbf{x}(2)), \dots, \Phi(\mathbf{x}(t), \mathbf{x}(N))) \quad (10)$$

$$C_Y(i, j) = \Phi(\mathbf{x}(i), \mathbf{x}(j)) \quad i, j \in [1, N] \quad (11)$$

Here N represents the dimension of Y . As Y describes a sequence of data in time series, the elements of Y could be sampled long while ago. When training the model to identify the hyper-parameters using (6) and making new predictions using (9), it can be seen that the computation complexity is $O(N^3)$. So it could be quite computationally expensive when large training data is used.

3. A variant of Gaussian Process

120 As the time series model ignores the exogenous inputs of weather information, the wind power output can be viewed as a system with time varying characteristic. The distribution of the newly sampled data may differ from that of the former one. Thus it is not reasonable to assume all the historical power data follow one joint Gaussian distribution. Besides, the computation complex-
 125 ity is unbearable when the number of available data increases. Under such a circumstance, a variant GP is proposed aiming at solving these two issues.

3.1. Method description

While the standard GP assumes that all the historical data Y follow one joint Gaussian distribution with the new one to predict $y(t)$, the proposed method
 130 considers the system to be stable only within a short period $M*\Delta T$ where ΔT is the sampling interval, and M is a positive integer. Denote $Y(t) = (y(t-1), y(t-2), \dots, y(t-M))^T$ describing the effective data window, then $Y(t)$ instead of Y ($Y(t) \subset Y$) is employed for the output prediction as shown in (12) and (13). As the number of used data is reduced in each prediction, the computation
 135 complexity will consequently be saved. Moreover, the prediction accuracy could still be satisfactory because only the highly correlated data are employed and potential extra error introduced by data sampled long while ago is removed.

$$\hat{y}(t) = B_\theta(t)C_\theta^{-1}(t)Y(t) = B_\theta(t)C_\theta^{-1}(t) \begin{pmatrix} y(t-1) \\ y(t-2) \\ \vdots \\ y(t-M) \end{pmatrix} \quad (12)$$

$$\sigma_{y_t}^2 = A(t) - B_\theta(t)C_\theta^{-1}(t)B_\theta^T(t) \quad (13)$$

$$\mathbf{x}(t-i) = (y(t-i-1), y(t-i-2), \dots, y(t-i-L)) \quad (14)$$

(12) and (13) show the inference procedure of the proposed GP. Here every intermediate variable such as $B_\theta(t)$, $C_\theta^{-1}(t)$ and $Y(t)$ changes according to time

instants in the following mechanisms. The point to predict at time t decides the local data within effective window $dY(t)$. Consequently, $C_\theta^{-1}(t)$, the self-covariance matrix of $Y(t)$, and $B_\theta(t)$, the cross covariance vector between new prediction $\hat{y}(t)$ and the available data $Y(t)$, are time dependant accordingly. These variables are quite different from the static covariance matrix C_Y and contribution data Y in (9) and could be expressed in (15) and (16). Here $\Phi(\cdot)$ refers to the covariance function as illustrated in Section 2.

$$B_\theta(t) = (\Phi(\mathbf{x}(t), \mathbf{x}(t-1)), \dots, \Phi(\mathbf{x}(t), \mathbf{x}(t-M))) \quad (15)$$

$$C_\theta(t) = \begin{pmatrix} \Phi(\mathbf{x}(t-1), \mathbf{x}(t-1)), & \dots & \Phi(\mathbf{x}(t-1), \mathbf{x}(t-M)) \\ \vdots & \ddots & \vdots \\ \Phi(\mathbf{x}(t-M), \mathbf{x}(t-1)) & \dots & \Phi(\mathbf{x}(t-M), \mathbf{x}(t-M)) \end{pmatrix} \quad (16)$$

Similarly, the learning procedure of proposed GP differs from the standard one as well. In the learning procedure of a standard GP as shown in (6), the likelihood of the available data following joint Gaussian distribution is maximized, thus the optimal hyperparameters are obtained. In this proposed method, the same number of data are employed. After the hyperparameters are initialized, each of the training dataset is predicted with its individual local data using (12), then the sum of square errors is calculated and minimized with nonlinear optimization techniques. For N samples of training data, the objective function could be expressed as (17) where J refers to the cost function and M defines the length of effective window.

$$\begin{aligned} \theta^* &= \arg \min_{\theta} J = \arg \min_{\theta} \sum_{k=M+1}^N (\hat{y}_k - y_k)^2 \\ &= \arg \min_{\theta} \sum_{k=M+1}^N (B_\theta(k)C_\theta(k)^{-1}Y(k) - y_k)^2 \end{aligned} \quad (17)$$

The proposed prediction method works like a moving window for a sequence of consecutive data. The data points are estimated one by one with a window of

Table 1: The computation time and complexity of standard GP and the proposed method ($m \ll N$)

Computation complexity	Learning	Inference	Uncertainty
Standard GP	$O(N^3)$	$O(N^3)$	$O(N^3)$
Proposed GP	$O(N * M^3)$	$O(M^3)$	$O(M^3)$

corresponding local data. As time moves forward, the effective window scrolls
 160 with time as well. In such circumstances, the ‘moving window’ technique is
 employed in the proposed method.

3.2. Computation complexity analysis

From the description illustrated in the above section, it can be seen that the
 proposed GP regression method is a compromise between a global method and
 165 a local method. It shows a global property when it employs all the data to train
 the model and find the optimal hyperparameters which guarantees the accuracy
 of the proposed method. On the other hand, it utilizes only local data in the
 inference procedure which removes the unnecessary effect brought by irrelevant
 data sampled long while ago and also reduces the computation complexity at
 170 the same time.

Table 1 shows the computation complexity comparison of the standard method
 and the proposed method. The complexity of the inference and the marginal
 likelihood function is determined by the size of covariance matrix because of the
 inversion operation. In the proposed method, the matrix dimension is reduced
 175 from N to M , so the computation complexity of the point inference equation
 (12), and the variance uncertainty estimation (13), is reduced from $O(N^3)$ to
 $O(M^3)$. In minimizing SSE (17), the point inference is implemented $(N - M)$
 times in the learning process, so the complexity involved is $O(N * M^3)$ which
 is still much smaller than that of standard GP $O(N^3)$, due to the fact that
 180 $m \ll N$.

3.3. Model Consistency

Consistency is a desirable property for supervised learning techniques. As more and more data are obtained, it would be expected that the predictions could converge to the true underlying predictive distribution. Hence, if the proposed method is consistent, the estimation $\hat{f}(\mathbf{x}_t)$ should converge to the real value $f(\mathbf{x}_t)$ as shown in (18) when the number of training data approaches infinite.

$$\hat{y}_t = \hat{f}(\mathbf{x}_t) \rightarrow f(\mathbf{x}_t) \quad (18)$$

To predict output y_t at time t with the proposed method, suppose there are N data between $t - T$ and t for model training, and M data in the local data window between $t - T_W$ and t for new prediction ($N \gg M$, and $T \gg T_W$). Thus T_W could be translated as the width of the effective data window. Under these circumstances, the consistency of the proposed model could be described in Theorem 1.

Theorem 1. Consider the proposed variant GP described in equations (12)-(17). Suppose as the total data points $N \rightarrow \infty$ and the effective window width $T_W \rightarrow 0$, the number $M(N, T_W)$ of data points in the local window goes to infinite. Further assume that the unknown $f(\cdot)$ is continuous in the neighbourhood of x_t and thus has a bounded first derivative, then in probability, the estimation $\hat{y}_t = \hat{f}(x_t)$ by the proposed method converges to the underlying function output $f(x_t)$.

Proof. First, consider two matrices

$$(1, 1, \dots, 1) \in R^{1 \times n}, \quad \begin{pmatrix} 1 + \frac{1}{n} & 1 & \dots & 1 \\ 1 & 1 + \frac{1}{n} & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 + \frac{1}{n} \end{pmatrix} \in R^{n \times n} \quad (19)$$

Then the product of the first and the inversion of the second converges to a

special vector as shown in (20).

$$(1, 1, \dots, 1) \begin{pmatrix} 1 + \frac{1}{n} & 1 & \dots & 1 \\ 1 & 1 + \frac{1}{n} & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 + \frac{1}{n} \end{pmatrix}^{-1} \quad (20)$$

$$\rightarrow \frac{1}{n}(1, 1, \dots, 1)$$

Give θ an initial value as $\bar{\theta} = (1, \frac{1}{M}, 0, 0, \dots, 0, 0,)$ indicating that the coefficients of each dimension of state vector $(x_d(i) - x_d(j))|_{d=1}^D$ is set to be zero in the covariance function (5). Let $C_{\bar{\theta}}$ and $B_{\bar{\theta}}$ denote the covariance matrix C_{θ} and B_{θ} respectively when $\theta = \bar{\theta}$ and y denote the measured value of the output. By substituting $\bar{\theta}$ into the proposed method (12), it follows that

$$\hat{y}_t|_{\bar{\theta}} = B_{\bar{\theta}} C_{\bar{\theta}}^{-1} Y_t = (1, 1, \dots, 1) \times$$

$$\begin{pmatrix} 1 + \frac{1}{M} & 1 & \dots & 1 \\ 1 & 1 + \frac{1}{M} & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 + \frac{1}{M} \end{pmatrix}^{-1} \begin{pmatrix} y(t-1) \\ y(t-2) \\ \vdots \\ y(t-M) \end{pmatrix} \quad (21)$$

$$\rightarrow \frac{1}{M} \sum_{i=1}^M y(t-i)$$

$$\rightarrow \frac{1}{M} \left(\sum_{i=1}^M f(\mathbf{x}_{t-i}) + \sum_{i=1}^M v_{t-i} \right)$$

The second term refers to the sum of measurement noise therefore it converges to zero by the law of large numbers ($M \rightarrow \infty$). For the first term, note

$$f(\mathbf{x}_{t-i}) = f(\mathbf{x}_t) + \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} (\mathbf{x}_t - \mathbf{x}_{t-i}) \quad (22)$$

where \mathbf{x}_{t-i} is a sample within the effective window and \mathbf{x} is some point between \mathbf{x}_{t-i} and \mathbf{x}_t . As $f(\cdot)$ is continuous in the neighbourhood of \mathbf{x}_t , the absolute value of the differential is finite which can be described as $|\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}| \leq \alpha$. It could be derived from (22) that

$$|f(\mathbf{x}_{t-i}) - f(\mathbf{x}_t)| \leq \alpha |\mathbf{x}_t - \mathbf{x}_{t-i}| \quad (23)$$

Within the effective time window $WT \rightarrow 0$, we have $|\mathbf{x}_t - \mathbf{x}_{t-i}| \rightarrow 0$, so $f(\mathbf{x}_{t-i}) \rightarrow f(\mathbf{x}_t)$. Consequently, the convergence result shown in (24) can be derived based on (21).

$$\hat{y}_t|_{\bar{\theta}} \rightarrow f(\mathbf{x}_t) \quad (24)$$

Now, consider the training data y_k ($k \in [M+1, N]$) in learning process, let t in equation (24) is substituted with k , then

$$\begin{aligned} J(\bar{\theta}) &= \sum_{k=M+1}^N (\hat{y}_k - y_k)^2|_{\bar{\theta}} = \sum_{k=M+1}^N (\hat{y}_k - f(\mathbf{x}_k))^2|_{\bar{\theta}} \\ &+ \sum_{M+1}^N 2v(k)(\hat{y}_k - f(\mathbf{x}_k))|_{\bar{\theta}} + \sum_{M+1}^N v(k)^2 \end{aligned} \quad (25)$$

The last term is independent of the hyper-parameters θ and the second term converges to zero because of i.i.d noise of zero mean. Combined with the above equation and the fact that $J(\theta^*) \leq J(\bar{\theta})$, then as $N \rightarrow \infty$, we have

$$\sum_{k=M+1}^N (\hat{y}_k - f(\mathbf{x}_k))^2|_{\theta^*} \leq \sum_{k=M+1}^N (\hat{y}_k - f(\mathbf{x}_k))^2|_{\bar{\theta}} \quad (26)$$

Considering (24), the above equation implies

$$(\hat{y}_k - f(\mathbf{x}_k))^2|_{\theta^*} \rightarrow 0 \quad (27)$$

This shows $\hat{y}_k \rightarrow f(\mathbf{x}_k)$ for each $k = M+1, \dots, N$. The convergence property at the testing data can be given following a similar approach.

□

3.4. Multi-step prediction

In wind power forecasting, wind power generation data are normally sampled with an interval of 10 or 15 minutes. In order to achieve several hours ahead prediction, multi step prediction is required. For iterative multi step prediction, $y(t+1)$ is first estimated with the proposed method illustrated above and then the new estimation is used to construct the new state vector $\mathbf{x}(t+2)$ and predict $y(t+2)$ with (12) again. And similarly, the prediction propagates from $y(t+1)$ to $y(t+Q)$ step by step where Q is an positive integer number. In such a manner,

the future output forecasting employs the same model at every step with fixed hyperparameters θ and covariance matrix C_Y , so there is no need to train the model separately to adapt to different steps prediction. Iterative multi step forecasting is a computationally efficient way in comparison with those direct ones [40].

4. Nonlinear optimization technique

In the above Section, the proposed model is shown to have simpler computation complexity and learning consistency. As can be seen, the nonlinear optimization problem in developing the proposed model can be completely different from the standard GP: the former one uses minimizing least square method, while the later one utilizes maximizing marginal likelihood method. Considering that the optimization problem has multiple local minima and conventional optimization methods may be less effective, in this section, the meta-heuristic methods for optimizing the nonlinear fitness function in (17) will be investigated.

There are many meta-heuristic optimization methods such as Genetic Algorithm (GA) and Particle Swarm optimization (PSO), etc. which are inspired by nature. Based on the Darwin's Theory Of Evolution, the process of GA is controlled by two parameters: crossover rate and mutation rate. Similarly, PSO imitates the foraging behaviour of birds and uses inertia weight, social and cognitive to adjust the process [41]. The choice of those parameters can have a large impact on optimization performance. Teaching-learning based optimization, proposed by Rao et.al [33] in 2011 is a new method to remove the tedious procedure in selecting those parameters, i.e. apart from several common parameters like population size and evolution generation, there is no specific algorithm parameters. Particles mutation depend solely on the statistics information of the whole population and solutions interactions. The procedure of TLBO has two phases and is illustrated as follows.

4.1. Teaching Phase

265 In each iteration the best solution in the particles will be first selected and called a 'teacher' after comparing all the fitness function values of the whole population, and all the other particles are called 'students'. Hence it makes sense for all the students to move towards the teacher to learn and improve. The mean of all the particles is calculated to reflect the average level of the students.

270 In order to reflect the general studying ability of the class and differ the ability of different students, two kinds of random values r_i and m are introduced in the following equations to construct the moving direction of each particle in every iteration. As i reflects the iteration number, r_i stays the same in every iteration and $r2$ changes for every student. These two random variables enhance the

275 exploitation ability of this algorithm.

$$DM_i = r_i \times (T_i - T_F Mean_i) \quad (28)$$

$$T_F = round(1 + r2) = round(1 + rand(0,1)) \quad (29)$$

Here $Mean_i$ denotes the mean of all the solutions of i -th iteration while DM_i is the moving direction to update those solutions. T_i denotes the selected teacher, and T_F is called the teaching factor. T_F can be either 1 or 2. The new positions

280 of these students can be updated as follows

$$\theta_i^{new} = \theta_i^{old} + DM_i \quad (30)$$

where θ_i^{new} and θ_i^{old} denote the old and new status of population. The fresh learners will compete with its predecessor and replace if a better fitness value is achieved.

4.2. Learning Phase

285 In a class, the interaction between students has an important impact for their growth. Similarly, such effect could be reflected in the learner phase of each iteration, when each of the population learn from a random student and

update himself or herself accordingly. The learning phase could be expressed as follows

$$\theta_{ij}^{new} = \begin{cases} \theta_{ij}^{old} + r3(\theta_{ik} - \theta_{ij}) & \text{if } f(\theta_{ik}) < f(\theta_{ij}) \\ \theta_{ij}^{old} + r3(\theta_{ij} - \theta_{ik}) & \text{if } f(\theta_{ij}) < f(\theta_{ik}) \end{cases} \quad (31)$$

290 Here i still refers to the iteration number. The j^{th} learner θ_{ij} and k^{th} learner θ_{ik} are randomly selected from the population, compared with each other, and finally updated accordingly. $r3$ is a random value represent the extend learners learn from each other and it changes for different learners. It should be noted that the new solution will have to compete with the old one. Only if the fitness
 295 gets better, the new value will get accepted, otherwise rejected, just similar to that of the teaching phase.

Besides the common initialization parameters such as the population size and the termination criteria, none other parameters has been introduced into the optimization process. Hence, this algorithm overcomes several technical
 300 problems. Moreover, the consistency of the algorithms has been proved as well with some well known benchmarks in [33] showing the efficient computation process. Hence, TLBO is selected as the optimization technique for the proposed variant GP regression method.

The proposed TLBO based variant Gaussian Process could be described with
 305 Fig.1. After maximum iteration number is reached, the optimization process terminates.

5. Case studies and prediction results

In the Republic of Ireland and Northern Ireland, wind power has been set as the main renewable resource due to the highly available wind resource.

310 Fig.2 displays the installed wind power on a county by county basis in Ireland. In this section, the wind generation of the whole island and that of a small wind farm in Donegal are both predicted representing forecasting examples of different scales.

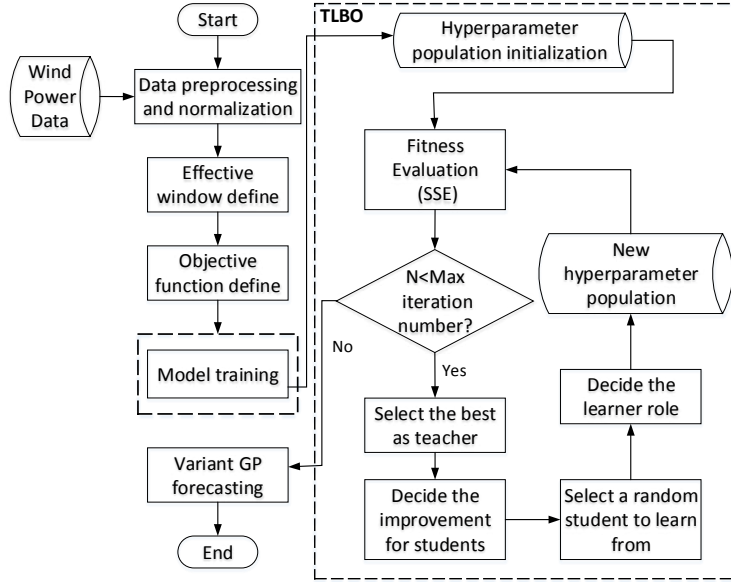


Figure 1: The flow chart of variant GP based on TLBO

5.1. Whole island wind power forecasting

315 Power forecasting for Ireland, which includes the Republic of Ireland and Northern Ireland is important as the generation mix is significant in terms of wind penetration. Currently, the governments in the British Isles and France have focused on increased cooperation between the different regions to increase co-operation for grid balancing, wind integration, security of energy supply and

320 reduce greenhouse gas emissions in order to meet European Union (EU) energy targets. The France, United Kingdom and Ireland is referred to as the FUI region in the wider EU energy balancing areas. The Single wholesale Electricity Market (SEM), which includes the Republic of Ireland and Northern Ireland and the British Trading and Tarriff Arrangment (BETTA), which includes the

325 regions of England, Scotland and Wales are trying in essence to improve internally balancing to better facilitate the planned increases in onshore and offshore wind power [42]. Under such circumstances, the wind power forecasting of the

first normalized so they are in a range of $(-1, 1)$ with mean value of zero and then the predictions were denormalized to reflect the corresponding estimations. According to some trial-and-error experiments, the length of the state vector was set as $L=10$, generating 12 parameters in all to tune, and $M = 14$, suggesting the width of local window. TLBO was employed to identify those hyper-parameters and the global optima could be approximately approached with multi simulation and proper setting of initial points. The validation results are shown in Fig.3 which shows that the error is below 10% of the actual power measurements, suggesting a good forecasting performance.

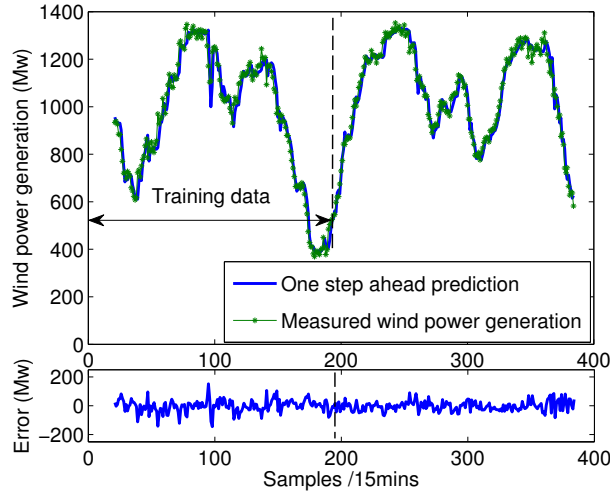


Figure 3: One step ahead prediction result of the proposed method for the whole island

In comparison with the standard GP and the well known persistence model in wind power forecasting, the proposed method proves its effectiveness again with smaller normalized RMSE and MAE over multi step prediction as shown in Fig.4 and Fig.5 respectively.

TLBO is adopted as the optimization method in comparison with standard particle swarm optimization (PSO) [43] with $c1=1$, $c2=1.5$ and genetic algorithm (GA) [44] with $Cr=0.8$, $Mu=0.2$. To fairly compare the performances, the well-known function evaluations (FES) criterion is employed. It should be

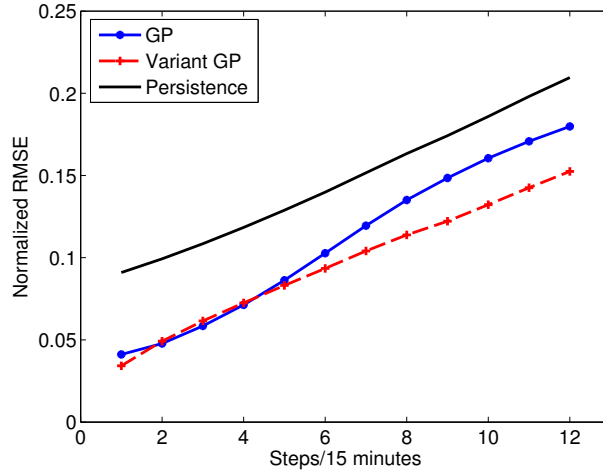


Figure 4: The normalized RMSE of different models for the whole island

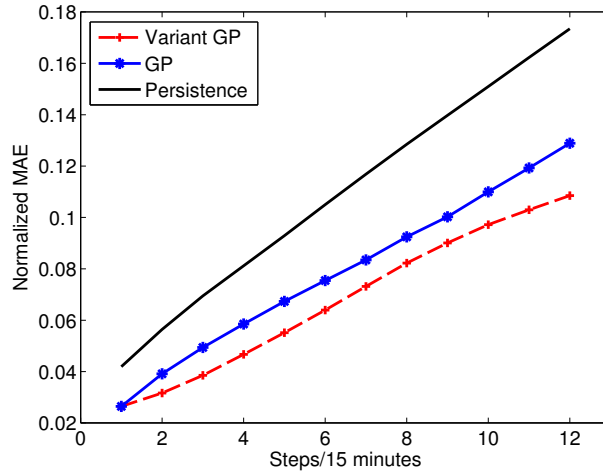


Figure 5: The normalized MAE of different models for the whole island

noted that the two phases TLBO algorithm has doubled the FES within the same evolutionary iteration of GA and PSO. The FES of algorithms test is 4500
 360 through which the algorithms converge in the test. The number of particles (Np) in PSO, GA and TLBO is set as 20, 30 and 50 respectively. While the number of iteration (IterMax) are set as 225, 150 and 90 for PSO and GA, it is set as 112, 75 and 45 for the TLBO at different particle numbers. To eliminate

the randomness, 10 independent tests are implemented and listed in the Table

365 2

As the result shown in Table 2, TLBO achieves the best results comparing with two counterparts in all population settings, where the best result is achieved with the configuration of $Np = 50$. Moreover, the average optimization process is shown in Fig.6. It could be easily observed that TLBO rapidly converged to a relatively low training errors in 1000 FES, significant outperforming PSO and GA. This optimization process suggests the advantage of TLBO algorithm in seeking the optimal solution when the hyperparameters are set within a large range which is very useful when there is no prior information about where the optimal solution would be located. Note that besides the original TLBO, some TLBO variants, such as mTLBO [45], weighted TLBO [46] and SL-TLBO[35] could further be utilized for model refinement, which will not be addressed in this paper due to the space limitation.

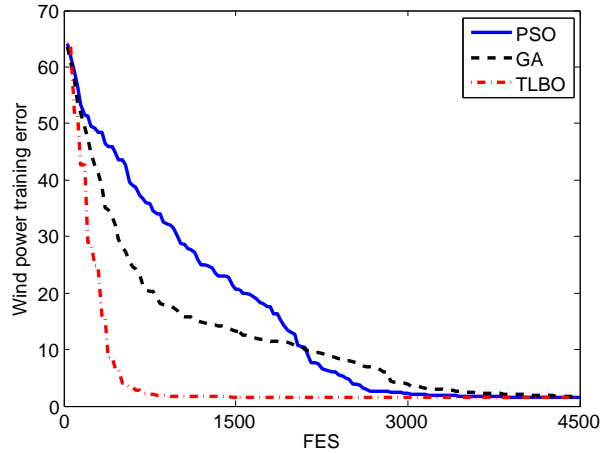


Figure 6: The average optimization process of TLBO compared with PSO and GA

5.2. Small wind farm forecasting

In this part, the proposed method and the uncertainty propagation analysis are applied to a small wind farm located in the Donegal area of Republic of Ireland which is labelled as ‘A’ in the top left corner of Fig.2. With the influence

380

Table 2: Optimisation results of GA, PSO and TLBO methods

Population size	<i>GA</i>			<i>PSO</i>			<i>TLBO</i>		
	Best	Mean	Worst	Best	Mean	Worst	Best	Mean	Worst
$N_p=20$	1.6991	2.2130	2.9173	1.5645	1.6530	1.7101	1.5574	1.6429	1.7702
$N_p=30$	1.7088	1.9720	2.6902	1.5577	1.6022	1.6738	1.5262	1.5698	1.6555
$N_p=50$	1.6026	1.6972	1.8573	1.5717	1.6532	1.7586	1.5138	1.5606	1.6329

of Atlantic sea wind and lake-hill breeze, the wind farm outputs show high variability and great intermittency. Further, the generation is more unpredictable in this wind farm due to its small capacity. However, its large noise and strong discontinuity will make the proposed model more convincing in comparison with other less complex systems.

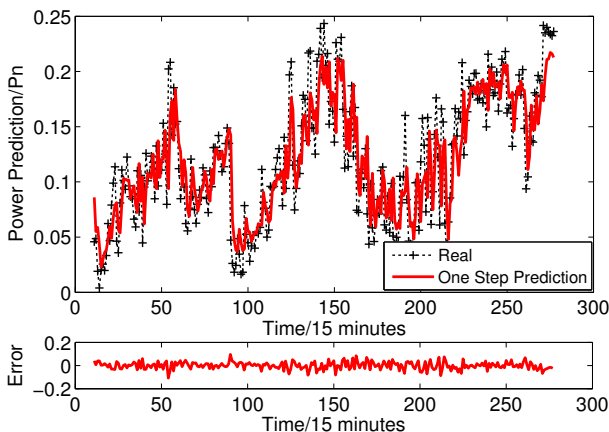


Figure 7: One step ahead prediction result of the proposed method for a wind farm

Data was collected on the last week of June of 2004 in a time interval of 15 minutes to predict the output of the first three days of July. The data was first normalized with the wind farm capacity, and then the squared exponential covariance function was employed with the mean function set as zero. With the proposed model, employing TLBO as the optimization method to minimize the cost function (17), the forecasting results are shown in Fig.7. . It can be seen that the regression error is a little bit bigger than that of the whole island due to

the fact that these data are much smaller and thus contains higher percentage
395 of noise.

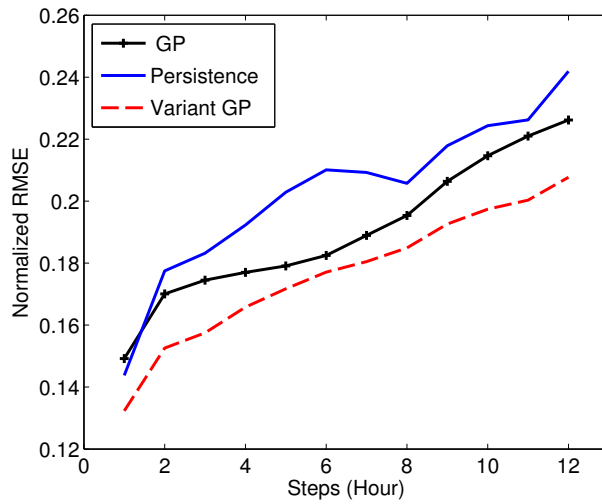


Figure 8: The normalized RMSE comparison of three forecasting models

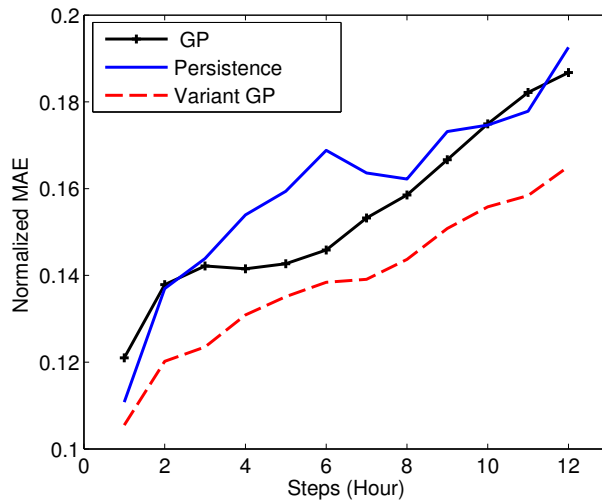


Figure 9: The normalized MAE comparison of three forecasting models

Fig.8 and Fig.9 fig compares the prediction error of the three different models. It shows that the proposed temporally local Gaussian Process (TLGP)

outperforms the other two with better accuracy from 1 hour ahead to 12 hours ahead.

400 Further, the optimization process over the forecasting procedure for this small wind farm is similar to the whole island in Fig.6. It is obvious that TLBO shows its advantages for searching the optimal solutions within a large search range.

6. Conclusions

405 In this paper, a variant GP modelling method which integrates local property into the global GP has been proposed. The method is specially designed for time series models, and could be adapted to time varying systems with greatly reduced computation complexity. Further, the consistency of the proposed model has been proved in this paper for cases where a large number of
410 samples are available. To train the model, a new optimization method namely TLBO is applied in the learning process to obtain the optimal solution. Experimental results show that TLBO has a better global exploitation ability and a faster convergence rate. The proposed methods have been applied to short-term forecasting of wind power generation for both the whole Ireland and for a small
415 wind farm. The case studies confirm the effectiveness of the proposed method for different scales of wind generation or different degrees of noise influence.

As Gaussian Process provides additional uncertainty information beside the mean value prediction, the method proposed in this paper, which is based on GP, could be further employed for probabilistic wind power forecasting, which would
420 benefit the system operation and scheduling at different time horizons. The key points remained to be solved would be the law of uncertainty propagation with multi-step forecasting, the calculation of the confidence level for the new predictions, and the evaluation of the accuracy and stability of uncertainty estimation. In the future research, those aspects would be studied and the
425 results should be compared with standard GP for further discussion. Further, the variants of TLBO such as mTLBO, weighted TLBO and SL-TLBO will

be investigated and applied on this problem. The exploration ability will be compared and analysed. This kind of research would provide more information for the operators and thus benefit both the social and economic effects.

430 **Acknowledgement**

This work is partially funded by Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/L001063/1 and EP/G042594/1. Many thanks to Chinese Scholarship Council (CSC) for UK-China Science Bridge Project for their sponsorship to Juan Yan during her PhD study at Queen's
435 University Belfast. Moreover, Dr Foley gratefully acknowledges EirGrid for use of datasets from the SEM.

References

- [1] G. Giebel, R. Brownsword, G. Kariniotakis, M. Denhard, C. Draxl, The state-of-the-art in short-term prediction of wind power: A literature
440 overview, Tech. rep., ANEMOS. plus (2011).
- [2] A. M. Foley, P. G. Leahy, A. Marvuglia, E. J. McKeogh, Current methods and advances in forecasting of wind power generation, *Renewable Energy* 37 (1) (2012) 1–8.
- [3] E. Erdem, J. Shi, Arma based approaches for forecasting the tuple of wind
445 speed and direction, *Applied Energy* 88 (4) (2011) 1405–1414.
- [4] G. Sideratos, N. D. Hatziargyriou, Probabilistic wind power forecasting using radial basis function neural networks, *Power Systems, IEEE Transactions on* 27 (4) (2012) 1788–1796.
- [5] T. Barbounis, J. Theocharis, Locally recurrent neural networks for long-
450 term wind speed and power prediction, *Neurocomputing* 69 (4) (2006) 466–496.

- [6] O. Kramer, F. Gieseke, B. Satzger, Wind energy prediction and monitoring with neural computation, *Neurocomputing* 109 (2013) 84–93.
- [7] M. Mohandes, S. Rehman, S. Rahman, Estimation of wind speed profile using adaptive neuro-fuzzy inference system (anfis), *Applied Energy* 88 (11) (2011) 4024–4032.
- [8] M. Poncela, P. Poncela, J. R. Perán, Automatic tuning of kalman filters by maximum likelihood methods for wind energy forecasting, *Applied Energy* 108 (2013) 349–362.
- [9] P. Kou, F. Gao, X. Guan, Sparse online warped gaussian process for wind power probabilistic forecasting, *Applied Energy* 108 (2013) 410–428.
- [10] N. Chen, Z. Qian, I. Nabney, X. Meng, Wind power forecasts using gaussian processes and numerical weather prediction, *Power Systems, IEEE Transactions on* 29 (2) (2014) 656–665.
- [11] D. Lee, R. Baldick, Short-term wind power ensemble prediction based on gaussian processes and neural networks, *Smart Grid, IEEE Transactions on* 5 (1) (2014) 501–510.
- [12] R. V. Mises, *Mathematical Theory of Probability and Statistics*, Academic Press, New York, 1964.
- [13] A. O’Hagan, J. Kingman, Curve fitting and optimal design for prediction, *Journal of the Royal Statistical Society. Series B (Methodological)* (1978) 1–42.
- [14] C. E. Rasmussen, *Gaussian Processes for machine learning*, The MIT Press, Cambridge, MA, 2006.
- [15] D. J. MacKay, Introduction to monte carlo methods, in: *Learning in graphical models*, Springer, 1998, pp. 175–204.
- [16] C. K. Williams, Regression with gaussian processes, in: *Mathematics of Neural Networks*, Springer, 1997, pp. 378–382.

- [17] X. Hong, J. Gao, X. Jiang, C. J. Harris, Fast identification algorithms for gaussian process model, *Neurocomputing* 133 (2014) 25–31.
- [18] M. N. Gibbs, D. J. MacKay, Variational gaussian process classifiers, *IEEE Transactions on Neural Networks* 11 (6) (2000) 1458–1464.
- [19] L. Wang, C. A. Leckie, Improved gaussian process classification via feature space rotation, *Neurocomputing* 83 (2012) 89–97.
- [20] E.-W. Bai, Local prediction error adjusted gaussian process for nonlinear non-parametric system identification, in: *System Identification*, Vol. 16, 2012, pp. 101–106.
- [21] D. Gu, H. Hu, Spatial gaussian process regression with mobile sensor networks, *Neural Networks and Learning Systems*, *IEEE Transactions on* 23 (8) (2012) 1279–1290.
- [22] S. Kim, J. Kim, Occupancy mapping and surface reconstruction using local gaussian processes with kinect sensors, *Cybernetics*, *IEEE Transactions on* 43 (5) (2013) 1335–1346.
- [23] X. Zhao, Y. Fu, Y. Liu, Human motion tracking by temporal-spatial local gaussian process experts, *Image Processing*, *IEEE Transactions on* 20 (4) (2011) 1141–1151.
- [24] C. Wu, Y. Wang, H. R. Karimi, A robust aerial image registration method using gaussian mixture models, *Neurocomputing* 144 (2014) 546–552.
- [25] J. Zhu, S. Sun, Sparse gaussian processes with manifold-preserving graph reduction, *Neurocomputing* 138 (2014) 99–105.
- [26] J. Yu, Semiconductor manufacturing process monitoring using gaussian mixture model and bayesian method with local and nonlocal information, *Semiconductor Manufacturing*, *IEEE Transactions on* 25 (3) (2012) 480–493.

- 505 [27] S. Faul, G. Gregorcic, G. Boylan, W. Marnane, G. Lightbody, S. Connolly, Gaussian process modeling of eeg for the detection of neonatal seizures, Biomedical Engineering, IEEE Transactions on 54 (12) (2007) 2151–2162.
- [28] L. Pasolli, F. Melgani, E. Blanzieri, Gaussian process regression for estimating chlorophyll concentration in subsurface waters from remote sensing
510 data, Geoscience and Remote Sensing Letters, IEEE 7 (3) (2010) 464–468.
- [29] J. Quiñonero-Candela, C. E. Rasmussen, A unifying view of sparse approximate gaussian process regression, The Journal of Machine Learning Research 6 (2005) 1939–1959.
- [30] L. Csató, M. Opper, Sparse on-line gaussian processes, Neural computation
515 14 (3) (2002) 641–668.
- [31] E. Meeds, S. Osindero, An alternative infinite mixture of gaussian process experts, Advances in Neural Information Processing Systems 18 (2006) 883.
- [32] J. Yan, K. Li, E.-W. Bai, Prediction error adjusted gaussian process for short-term wind power forecasting, in: Intelligent Energy Systems (IWIES), 2013 IEEE International Workshop on, IEEE, 2013, pp. 173–178.
520
- [33] R. Rao, V. Savsani, D. Vakharia, Teaching–learning-based optimization: A novel method for constrained mechanical design optimization problems, Computer-Aided Design 43 (3) (2011) 303–315.
- [34] T. Niknam, R. Azizipanah-Abarghooee, J. Aghaei, A new modified
525 teaching-learning algorithm for reserve constrained dynamic economic dispatch, Power Systems, IEEE Transactions on 28 (2) (2013) 749–763.
- [35] Y. Zhile, L. Kang, N. Qun, X. Yusheng, A. FOLEY, A self-learning tlbo based dynamic economic/environmental dispatch considering multiple plug-in electric vehicle loads, Journal of Modern Power Systems and
530 Clean Energy 2 (4) (2014) 298–307.

- [36] Y. Xu, L. Wang, S.-y. Wang, M. Liu, An effective teaching–learning-based optimization algorithm for the flexible job-shop scheduling problem with fuzzy processing time, *Neurocomputing* 148 (2015) 260–268.
- [37] Z. Yang, K. Li, A. Foley, C. Zhang, A new self-learning tlbo algorithm for rbf neural modelling of batteries in electric vehicles, in: *Evolutionary Computation (CEC), 2014 IEEE Congress on, IEEE, 2014*, pp. 2685–2691. 535
- [38] C. E. Rasmussen, *Gaussian processes for machine learning*, in: *Adaptive Computation and Machine Learning*, Citeseer, 2006.
- [39] M. N. Gibbs, *Bayesian gaussian processes for regression and classification*, Ph.D. thesis, Citeseer (1997). 540
- [40] A. Girard, C. E. Rasmussen, J. Quinero-Candela, R. Murray-Smith, Gaussian process priors with uncertain inputs? application to multiple-step ahead time series forecasting.
- [41] N. Ming, W. Can, X. Zhao, A review on applications of heuristic optimization algorithms for optimal power flow in modern power systems, *Journal of Modern Power Systems and Clean Energy* 2 (4) (2014) 289–297. 545
- [42] Open letter: Implementing the european electricity target model in great britain, Tech. rep., The Office of Gas and Electricity Markets (March 2012).
- [43] J. Kennedy, Particle swarm optimization, in: *Encyclopedia of Machine Learning*, Springer, 2010, pp. 760–766. 550
- [44] D. E. Goldberg, J. H. Holland, Genetic algorithms and machine learning, *Machine learning* 3 (2) (1988) 95–99.
- [45] S. C. Satapathy, A. Naik, Modified teaching–learning-based optimization algorithm for global numerical optimizationa comparative study, *Swarm and Evolutionary Computation* 16 (2014) 28–37. 555

- [46] S. C. Satapathy, A. Naik, K. Parvathi, Weighted teaching-learning-based optimization for global function optimization, *Applied Mathematics* 4 (03) (2013) 429.