# Analysis of Solar Generation and Weather Data in Smart Grid with Simultaneous Inference of Nonlinear Time Series

Yu Wang[†], Guanqun Cao[‡], Shiwen Mao[†], and R. M. Nelms[†]
[†]Department of Electrical and Computer Engineering, Auburn University, Auburn, AL
[‡]Department of Mathematics and Statistics, Auburn University, Auburn, AL
Email: yzw0032@tigermail.auburn.edu, gzc0009@auburn.edu, smao@ieee.org, nelmsrm@auburn.edu

***Abstract*—Smart Grid is an important component of Smart City, where more renewable power generation and better energy management is required. Forecast on renewable power generation, from sources such as solar and wind, is crucial for better energy management. However, the current forecast methods lack a comprehensive understanding of the natural processes, and are thus limited in precise prediction. In this paper, we introduce simultaneous inference to analyze the solar generation and weather data for better predictions. We first introduce a local linear model for nonlinear time series, and present the construction of the simultaneous confidence bands (SCB) of the time-varying coefficients, which provide more information on the dynamic properties of the model. We then use the simultaneous inference for solar intensity prediction using a real trace, where the superior performance of the proposed scheme is demonstrated over existing approaches.**

## I. Introduction

In recent years, as the development of the modern technologies in informatics, communication, control and computing, our living environment is becoming "smart." Smart Home and Smart City have gradually become part of our lives, and are no longer merely future concepts for the public. An important component of Smart City is the Smart Grid (SG), which is regarded as the next generation power grid to create a widely distributed energy generation and delivery network. The SG features the incorporation of power generation from renewable energy sources, especially solar and wind, which meanwhile requires a better energy management system in the SG [1].

Energy management in the SG has been studied in many previous works [1]–[3]. It is indicated in [3] that high efficient power management cannot be realized without a better forecast on the grid load and renewable power generation in the SG. The problem of grid load forecasting has been studied by many researchers with different techniques such as state space models [4], artificial neural networks and support vector machine (SVM) [5], and nonparametric functional time series analysis [3]. The prediction on the renewable energy generation in the SG has also attracted great interest. Predictions on solar and wind generation can be found in [6] and [7] using SVM regression and JPDF forecast respectively. Because of the weather dependence nature of the forecasting problem, statistical methods can be found in almost every related literature. On the other hand, the wide range of applications helps the improvement of the statistical theory on nonparametric analysis [8]–[10], and nonstationary time-series analysis [11], [12]. Recently, the authors of [13], [14] propose the method of constructing a simultaneous confidence band (SCB) for time-varying coefficients. These research advances improve the understanding of time-series, and provide better techniques in renewable energy generation prediction.

The previous work on predicting solar power generation provides acceptable results using SVM regression [6]. However, by simply trying different SVM kernels after some basic data processing statistically, it lacks a deep analysis of the solar power generation and weather data, and thus is limited in precise predictions of other data set. For example, the check of assumptions is missing on independence of variables and errors. Furthermore, the renewable energy generation is a function of weather variables, and is a stochastic process on nonlinear time series. Therefore, the relationship between the weather variables and the power generation should be analyzed over a long time for a comprehensive understanding of their dynamic relations. For example, a coefficient varying by time overall may stay constant for short periods. These drawbacks will limit the applications for predictions in other cases. Therefore, a method that can capture the dynamic property of the process would be highly desirable for better predictions on solar power generation in different scenarios.

Motivated by this observation, we introduce simultaneous inference of nonlinear time series proposed in [13] for understanding comprehensively the deep and dynamic relationship between renewable power generation process and the weather variable processes. The simultaneous inference is based on the SCB of time-varying coefficients in the local linear model, which we use for nonlinear time series analysis. It is based on the assumption of nonstationary processes for the error and weather variables, which matches the case of our problem where many weather variables are shown to have an obvious seasonal pattern, meaning the observations are not stationary assumed by many other forecast techniques. And the SCB shows the confidence bands of the coefficients over any length of time, which can be used to test if the coefficients are truly time-varying or not. This helps us to refine the model by omitting variables which that are not significant.

The main contribution of this paper is the introduction of

978-1-4673-7131-5/15/$31.00 ©2015 IEEE

the local linear model and SCB for analyzing the solar power generation as a nonlinear time series. By checking the dynamic properties of the coefficients from its SCB, we are able to achieve a more comprehensive understanding of the model, and based on this, we can further refine the model and use it for predicting the renewable energy generation. As an example, we apply it in predictions on daily solar power generation. This method has a wide range of applications, which is not limited to analyze and predict the solar energy generation. It can also be used for predictions in different time scales, from minutes to months, depending on different purposes, and in other cases, such as wind power generation prediction.

The remainder of this paper is organized as follows. We present the local linear model for nonlinear time-series analysis in Section II. The construction of SCB for time-varying coefficients with simulated results is introduced in Section III. We use the simultaneous inference for analyzing a trace of solar intensity and weather data in Section IV. Section V concludes this paper.

## II. Local Linear Model for Nonlinear Time Series

We consider the power generation from a renewable source as a continuous-time stochastic process $Y(t)$ and a function of the meteorological variables $\vec{X}(t)$, which is a continuous-time covariate process. It follows that

$$Y(t) = f(\vec{X}(t)), t \in \mathbb{R}.$$

To identify the function $f(\cdot)$, it is straightforward to try a linear model first, as

$$Y(t) = \vec{X}^T(t)\vec{\beta}(t) + \epsilon(t), t \in \mathbb{R}, \tag{1}$$

where $\vec{X}(t) = (1, X_1(t), ..., X_{p-1}(t))^T$ and $\vec{\beta}(t) = (\beta_1(t), ..., \beta_p(t))^T$ are both $p \times 1$ vectors, and $\epsilon(t)$ is the error at time $t$. To use this model, we need to predict the regression coefficients $\vec{\beta}(t)$ at each time $t$, so that we can calculate $Y(t)$ given the forecast on weather variables $\vec{X}(t)$. It is not difficult to obtain $\vec{\beta}(t)$ using parametric smoothing methods such as multiple linear regression. Although this model can indicate a certain level of interactions between the variables $\vec{X}(t)$ and the response $Y(t)$, it cannot be used to represent the real underlying process, especially when the method treat continuous-time series simply as discrete data points. As we will show in Section IV-D, the prediction with the linear model is not satisfactory, and cannot be used in practice.

From the linear model, we notice that the linearity is fairly strong between $\vec{X}(t)$ and $Y(t)$ within a short time period, i.e., several days or a week. If we take advantage of this property and consider the process as continuous-time, the model would be closer to the real process. For $t_i$ close to $t$, we can have $\vec{\beta}(t_i) \approx \vec{\beta}(t) + (t_i - t)\vec{\beta}'(t)$, where $\vec{\beta}'(t)$ is the derivative of $\vec{\beta}(t)$, and thus for any time $t_i$ close to $t$, we have the local linear model as [15]

$$Y(t_i) = \vec{X}^T(t_i)(\vec{\beta}(t) + (t_i - t)\vec{\beta}'(t)) + \epsilon(t_i), t_i \in t \pm h, \tag{2}$$

where the bandwidth $h$ is the size of the local neighborhood. This model divides the time series into periods and creates linear models using local data. This way, we treat the data as a continuous-time series, and exploit the strong correlations between close time periods in weather dependent systems.

### A. Local Linear Estimation

To identify the time-varying coefficients $\vec{\beta}(t)$, the least squares method for linear regression can be used. We also add some weights on the terms considering that contributions from different neighbors are different, which means a closer neighbor would have a stronger effect, while a further neighbor weaker effect. Usually a kernel function $K(\cdot)$ is assigned to each point, which is a symmetric density function defined on [-1,1] [15]. Here, we use a popular Epanechnikov kernel.

$$K(a) = \begin{cases} 3(1-a^2)/4, & \text{if } |a| \le 1 \\ 0, & \text{if } |a| > 1, \end{cases}$$

which decays fast for remote data point. We then have the following weighted least squares problem to solve.

$$\min : \sum_{t_i \in t \pm h} (Y(t_i) - \vec{X}^T(t_i)(\vec{\beta}(t) - (t_i - t)\vec{\beta}'(t)))K\left(\frac{t_i - t}{h}\right). \tag{3}$$

At each time $t$, we solve for the coefficients $\hat{\vec{\beta}}_h(t)$ and $\hat{\vec{\beta}}'_h(t)$ under the bandwidth $h$. Suppose the total number of observations is $n$, we can pick $t_i$ simply as $t_i = i/n, 1 \le i \le n$, and denote $Y(t_i)$ as $y_i$ and $\vec{X}(t_i)$ as $\vec{x}_i$. From [9], we can solve (3) by calculating the following matrices $\mathbf{S}_k(t)$ and $\mathbf{R}_k(t)$:

$$\mathbf{S}_k(t) = \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T \left(\frac{t_i - t}{h}\right)^k K\left(\frac{t_i - t}{h}\right)/(nh) \tag{4}$$

$$\mathbf{R}_k(t) = \sum_{i=1}^{n} \vec{x}_i y_i \left(\frac{t_i - t}{h}\right)^k K\left(\frac{t_i - t}{h}\right)/(nh), \tag{5}$$

where $k = 0, 1, 2, ....$ We then have

$$\begin{pmatrix} \hat{\vec{\beta}}_h(t) \\ h\hat{\vec{\beta}}'_h(t) \end{pmatrix} = \begin{pmatrix} \mathbf{S}_0(t) & \mathbf{S}_1^T(t) \\ \mathbf{S}_1(t) & \mathbf{S}_2(t) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{R}_0(t) \\ \mathbf{R}_1(t) \end{pmatrix}. \tag{6}$$

### B. Selection of Bandwidth

To solve problem (3) for the complete model using (4) to (6), we need to first fix bandwidth $h$. As discussed, $h$ is the bandwidth determining the size of data used to estimate for a local linear model at time $t$. If $h$ is too small, many useful points are not included for estimation, which may cause a larger error in the model; if it is too large, more remote points are included, which increases the computational complexity and reduces the smoothness of the model. Therefore, it is important to choose a proper $h$.

Popular bandwidth selection techniques can be found in [15], [16] for different applications. The techniques are different for constant bandwidth and variable bandwidth. For constant bandwidth selection considered in our model, we adopt the generalized cross-validation (GCV) technique [16], which is suitable for a wide range of applications.

Similar to multiple linear regression, the coefficients $\vec{\beta}$ are estimated from the observed data $\vec{Y}$ and $\vec{X}$. Thus, a square

hat matrix $\mathbf{H}(h)$ exists for $\hat{\vec{Y}} = \mathbf{H}(h)\vec{Y}$ [17], depending on the bandwidth $h$. Then we can choose the bandwidth $h$ by

$$\hat{h} = \arg\min \left\{ \frac{|\hat{\vec{Y}} - \vec{Y}|^2}{n(1 - \text{tr}\{\mathbf{H}(h)\}/n)^2} \right\}, \quad (7)$$

where, tr$(\cdot)$ is the trace of the matrix, and $n$ is the number of total observations.

## III. Simultaneous Confidence Band for Time-varying Coefficients

In this section, we introduce the basic conditions and construction of the simultaneous confidence band (SCB) method proposed in [13], and then discuss its implications to further understand the modeling and predicting for the power generation process from the renewable energy sources based on the weather data.

### A. Model Assumptions and Asymptotic Normality

Different from most current models for time series, the approach of SCB analysis assumes locally stationary processes for both $\vec{X}(t)$ and $\epsilon(t)$ [12]. The locally stationary process guarantees the stationary property for local time series, and is useful for local linear estimation. It actually belongs to a special class of non-stationary time series as

$$\vec{x}_i = \vec{G}(t_i, F_i), \quad \epsilon_i = H(t_i, F_i), \quad i = 1, 2, ..., n, \quad (8)$$

where $\vec{G}(t_i, F_i)$ and $H(t_i, F_i)$ are measurable functions well defined on $t_i \in [0, 1]$ , $F_i = (..., \xi_{i-1}, \xi_i)$ with $\{\xi_i\}_{i\in\mathbb{Z}}$ are independent and identically distributed (i.i.d.) random variables, and $\mathbb{E}(\epsilon_i|\vec{x}_i) = 0$. In our model for renewable power generation processes, we further assume that $\{\epsilon_i\}_{i\in\mathbb{Z}}$ are i.i.d. and dependent of $\{\xi_i\}_{i\in\mathbb{Z}}$.

Based on the above assumptions, the central limit theorem for $\hat{\vec{\beta}}(t)$ states that: supposing $nh \to \infty$ and $nh^7 \to 0$ [13], then for any fixed $t \in (0, 1)$,

$$(nh)^{1/2}\{\hat{\vec{\beta}}(t) - \vec{\beta}(t) - h^2\vec{\beta}''(t)\mu/2\} \to N\{0, \Sigma^2(t)\}, \quad (9)$$

where

$$\mu = \int_{\mathbb{R}} x^2 K(x)\, dx, \quad (10)$$

$$\Sigma(t) = (M^{-1}(t)\Lambda(t)M^{-1}(t))^{1/2}, \quad (11)$$

$$M(t) = \mathbb{E}(\vec{G}(t, F_0)\vec{G}(t, F_0)^T). \quad (12)$$

The covariance matrix $\Lambda(t)$ can be further approximated using techniques proposed in Section III-B.

### B. Simultaneous Confidence Band

Deriving from the central limit property and basic assumptions shown above, the $100(1-\alpha)\%$ asymptotic simultaneous confidence tube of $\vec{\beta}_C(t)$ can be constructed using the following formula:

$$\tilde{\beta}_{C,\tilde{h}}(t) + \hat{q}_{1-\alpha}\hat{\Sigma}_C(t)\mathcal{B}_s, \quad (13)$$

where $\tilde{\boldsymbol{\beta}}_{C,\tilde{h}}(t)$ is the bias corrected estimator defined in (14), $\mathcal{B}_s = \{\vec{z} \in \mathbb{R}^s : |\vec{z}| \le 1\}$ is the unit ball, and $s$ is the rank of a matrix $C_{p\times s}$, which we use for choosing different linear combinations of $\beta(t)$, and $\vec{\beta}_C(t) = C^T\vec{\beta}(t)$. To obtain the SCB, we simply take $s = 1$ in (13), and the SCB is constructed similarly to the confidence interval of the coefficients of the multiple linear regression: $\hat{\beta} \pm t_{\alpha/2,n-p}se(\hat{\beta})$, where se$(\hat{\beta})$ is the standard error of $\hat{\beta}$, and $t_{\alpha/2,n-p}$ is the upper $\alpha/2$ percentage point of the $t_{n-2}$ distribution [17].

**Algorithm 1:** Construction of SCB for Time-varying Coefficients

1. Find a proper bandwidth $\hat{h}$ from the GCV selector (7);
2. Let $\tilde{h} = 2\hat{h}$ and calculate $\tilde{\beta}_{C,\tilde{h}}(t)$ using (14) and (3);
3. Obtain the estimated $(1-\alpha)$th quantile $\hat{q}_{1-\alpha}$ via the bootstrap method;
4. Estimate $\hat{M}(t) = S_0(t^*)$ and $\hat{\Lambda}(t)$ by (**??**), and calculate $\hat{\Sigma}_C(t)$ according to (15);
5. Construct the $100(1-\alpha)\%$ SCB of $\vec{\beta}_C(t)$ using (13).

Similarly, the first term is the estimator of the time-varying coefficients corrected for bias by

$$\tilde{\beta}_{C,\tilde{h}}(t) = C^T\tilde{\beta}_{\tilde{h}}(t) = C^T\left(2\tilde{\vec{\beta}}_{\tilde{h}/\sqrt{2}}(t) - \hat{\vec{\beta}}_{\tilde{h}}(t)\right), \quad (14)$$

where $\tilde{\beta}_{\hat{h}}(t)$ can also be acquired by solving (3) using a corresponding kernel function $K^*(a) = 2\sqrt{2}K(\sqrt{2}a) - K(a)$ and an updated bandwidth $\tilde{h} = 2\hat{h}$ of the GCV selector $\hat{h}$.

The second term in (13) $\hat{q}_{1-\alpha}$ is actually the upper $\alpha/2$ percentage point of the normal distribution $N\{0, \Sigma^2(t)\}$ defined in (9), while the third term $\hat{\Sigma}_C(t)$ is the estimated standard error. The method of wild bootstrap is applied to obtain $\hat{q}_{1-\alpha}$. Firstly, generate a large number i.i.d. vectors $\vec{v}_1, \vec{v}_2, ...,\ N(0, I_s)$, where $\vec{v}_i \in \mathbb{R}^p$ and $I_s$ denotes the $s \times s$ identity matrix, and then calculate $q = sup_{0\le t\le 1}|\sum_{i=1}^n \vec{v}_i K^*((t_i - t)/\tilde{h})/(n\tilde{h})|$; repeat the previous step for a large number of times (say, 5000) to acquire the estimated $100(1-\alpha)\%$ quantile $\hat{q}_{1-\alpha}$ of $q$.

The estimate of the standard error in (13), $\hat{\Sigma}_C(t)$ is defined similarly as (11).

$$\hat{\Sigma}_C(t) = (C^T\hat{M}^{-1}(t)\hat{\Lambda}(t)\hat{M}^{-1}(t)C)^{1/2}. \quad (15)$$

We shall estimate $\hat{M}(t)$ and $\hat{\Lambda}(t)$ respectively. From the definition of $M(t)$ in (12), it can be estimated by $\hat{M}(t) = S_0(t^*)$, where $S_0(\cdot)$ is defined in (4), and $t^* = \max\{h, \min(t, 1-h)\}$. To obtain $\hat{\Lambda}(t)$, we first define two $p\times 1$ vectors $\vec{Z}_i = \vec{x}_i\hat{\epsilon}_i$ and $\vec{W}_i = \sum_{j=-m}^m \vec{Z}_{i+j}$, a matrix $\Omega_i = \vec{W}_i\vec{W}_i^T/(2m+1)$, and a function $g(t, i) = K((t_i - t)/\tau)/\sum_{k=1}^n K((t_k - t)$, where $m$ and $\tau$ can be simply chosen as $m = \lfloor n^{2/7} \rfloor$ and $\tau = n^{-1/7}$. Then $\hat{\Lambda}(t)$ can be calculated by $\hat{\Lambda}(t) = \sum_{i=1}^n g(t, i)\Omega_i$. This way, we are able to calculate the SCB using all the estimates. The above steps for constructing the SCB are summarized in Algorithm 1.

### C. Further Discussions

The SCB provides a dynamic and comprehensive view on $\vec{\beta}(t)$. In simple linear regression, the confidence interval

provides a measure of the overall quality of the regression line [17]. Similarly, the SCB illustrates the overall pattern of $\vec{\beta}(t)$ and thus the accuracy of the model. Confidence bands with smaller width implies a better model with smaller variability, while too wide confidence bands are limited in use. Note that the SCB is constructed under a complete analysis on the continuous-time assumption, which is not merely the connections of the point-wise confidence intervals on isolated time instances.

Further, the SCB can also be used to test whether the coefficients $\vec{\beta}(t)$ are truly time-varying or not. If a horizontal line is covered by the SCB of a $\beta_k(t)$, we accept the hypothesis that $\beta_k(t)$ is constant and not time-varying. Furthermore, in different cases, we can construct the SCB for different linear combinations of $\beta_k(t)$'s by setting a different matrix $C_{p\times s}$. For example, if we set $C_{p\times 1} = [1, 1, 0, ...0]$, we obtain the SCB of $\vec{\beta}_C(t) = C^T\vec{\beta}(t) = \beta_1(t) + \beta_2(t)$; if we set $C_{p\times 2} = [1, 0, 0, ...0; 0, 1, 0, ...0]$, the SCB of $\beta_1(t)$ and $\beta_2(t)$ becomes a tube at any time $t$. This is because when $s = 2$, the unit ball $\mathcal{B}_2$ turns to a unit circle from a unit interval. This provides us a convenient way to further test the model.

### D. Algorithm Performance for Simulated Processes

We now simulate a model with $\vec{X}(t)$ and $\epsilon(t)$ locally stationary processes discussed in Sec. III-A, and construct the $90\%$ and $95\%$ SCB respectively for a given model, to test the correctness by comparing with the true results.

We use the following local linear model with time-varying coefficient:

$$y_i = \beta_1(i/n) + \beta_2(i/n)x_i + \epsilon_i, \tag{16}$$

where $\beta_1(t) = \cos(2\pi t)/4$, and $\beta_2(t) = \exp\{-(t-1/2)^2\}/2$. Define $H(t_i, F_i) = (1/2)\sum_{j=0}^{\infty} a(t_i)^j\xi_{i-j}$, $\vec{G}(t_i, F_i) = (1; \sum_{j=0}^{\infty} b(t)^j\varepsilon_{i-j})$, where $\xi_k$ and $\varepsilon_l$ are i.i.d. $N(0, 1)$. Then $\vec{x}_i$ and $\epsilon_i$ can be generated using (8), for $i = 1, 2, ..., n$.

For the above setting, we generate 5000 samples of size 500, and for each sample SCB is constructed with bandwidths setting from 0.1 to 0.3 of step 0.025. We use 3000 and 5000 bootstrap samples to estimate $\hat{q}_{1-\alpha}$ for $\alpha = 0.1$ and $\alpha = 0.05$ to show the effect of the sample size on the results. The simulation results are shown in Table I, where the coverage rate and width of SCB for $\beta_2(t)$ with different bandwidths $h$ at $90\%$ and $95\%$ levels are listed. It shows that the coverage rate is close to the nominal level with most bandwidths. And the bandwidth selected by GCV is 0.22, which yield fairly good results. We also notice that $\hat{q}_{1-\alpha}$ are affected by the bootstrap sample size, and its value affects the width of SCB directly. Therefore, for practical application, a large size of bootstrap samples is very important. According to our numerical studies, at least 5000 samples are suggested.

## IV. Application to Solar Energy Generation

In this section, we apply the SCB analysis to modeling the solar power generation process and predicting the generation amount based on weather data. We also compare results with that of other methods.

TABLE I
THE COVERAGE PROBABILITIES OF SCB FOR $\beta_2(t)$ AND QUANTILES OF $\hat{q}$ AT NOMINAL LEVEL OF $90\%$ AND $95\%$

| | $\beta_2(t)$ | | 3000 Samples | | 5000 Samples | |
|---|---|---|---|---|---|---|
| bandwidth | $90\%$ | $95\%$ | $\hat{q}_{0.90}$ | $\hat{q}_{0.95}$ | $\hat{q}_{0.90}$ | $\hat{q}_{0.95}$ |
| 0.1 | 0.814 | 0.864 | 0.595 | 0.617 | 0.476 | 0.509 |
| 0.125 | 0.875 | 0.925 | 0.522 | 0.568 | 0.415 | 0.453 |
| 0.15 | 0.914 | 0.953 | 0.483 | 0.505 | 0.378 | 0.401 |
| 0.175 | 0.923 | 0.945 | 0.451 | 0.464 | 0.335 | 0.363 |
| 0.2 | 0.901 | 0.951 | 0.417 | 0.441 | 0.309 | 0.336 |
| 0.225 | 0.908 | 0.949 | 0.392 | 0.412 | 0.284 | 0.311 |
| 0.25 | 0.904 | 0.955 | 0.369 | 0.392 | 0.269 | 0.295 |
| 0.275 | 0.899 | 0.951 | 0.348 | 0.366 | 0.251 | 0.275 |
| 0.3 | 0.898 | 0.946 | 0.337 | 0.348 | 0.235 | 0.264 |

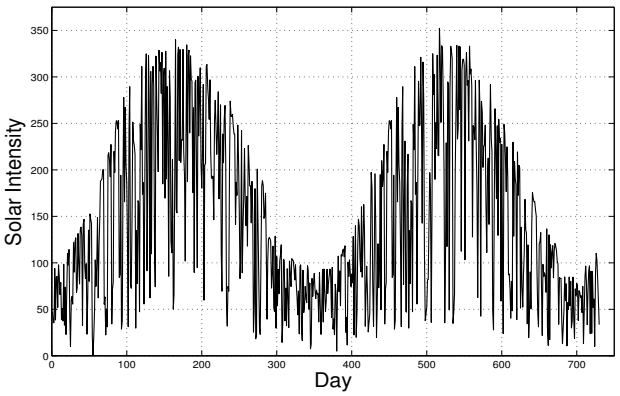


Fig. 1. Daily solar intensity for 2010 and 2011.

### A. Data Description

As an application, we consider the data from the UMASS Trace Repository [18], which records the solar power generation by solar intensity in $watts/m^2$, and the data of several weather metrics from January, 2010 to February, 2013. It recorded the weather data every five minutes. Many weather parameters were observed in details. We consider five main variables, including temperature, humidity, dew point, wind speed, and precipitation. The data has been used in [6] for a study of the statistical relationship between the weather variables and solar power generation. The paper also predicted solar power generation using multiple linear regression and Support Vector Machines regression. Our purpose is also to investigate the dynamic association between the weather variables and solar power generation, and to help better predict the solar power generation.

We plot the daily solar intensity for 2010 and 2011 in Fig. 1. An apparent seasonal pattern is shown with peak in summer and valley in winter. It would be helpful to consider the seasonal patterns for forecast. It is also interesting to see a similar pattern for daily observations, and similar patterns can also be seen for the other weather variables, such as temperature, humidity, and dew point (See [6]). Fig. 1 also

shows a strong correlation between two consecutive days, which means the solar generation process is not i.i.d. As discussed in Section III-A, we do not require the observations to be i.i.d. to construct the SCB.

### B. Prediction Model

As in Section (II), we use the following local linear model.

$$y_i = \beta_1 (i/n) + \sum_{p=2}^{5} \beta_p (i/n) x_{p,i} + \epsilon_i, \text{ for } i = 1, ..., n, \quad (17)$$

where $y_i$ is the solar intensity, $x_{p,i}$, $p = 2, 3, 4, 5$, represent the series of temperature in Fahrenheit, humidity in percentage, dew point in Fahrenheit, wind speed in miles per hour, and precipitation in inches, respectively. We use $n = 730$ observations of 2011 and 2012 for local linear regression and the model is represented in a daily pattern. Note our model and analysis can be built on any time scale. We take daily pattern for an application example here, where $\beta_1(\cdot)$ is the intercept and $\beta_p(\cdot)$ are the associated coefficients for $x_{p,i}$.

### C. Simultaneous Inference for Time-varying Coefficients

We now perform the SCB analysis. We center all the weather variables on their averages so that the intercept $\beta_1(\cdot)$ can be interpreted as the expected solar intensity. From GCV, we select the bandwidth $h = 0.25$. The $95\%$ SCB of the coefficients $\beta_p(\cdot)$ are shown from Fig. 2 to 7. In each figure, the middle thick solid curve is the estimated series for the variable; the upper and lower solid curves are the envelops for the simultaneous confidence band for each variable. From the SCB, we are able to test whether a coefficients is significantly associated with the solar intensity, which equals to test:

$$H_0 : \beta_p(t) = 0, \forall t \in [0, 1]; \text{v.s. } H_1 : \beta_p(t) \neq 0, \exists t \in [0, 1].$$

If the zero line is included in the SCB, we accept the hypothesis that the coefficient is not significant and could be omitted from the model; otherwise, we keep it in the model. We can also test whether the coefficients are constant, by attempting to include a constant horizontal line into the SCB. This is equal to testing:

$$H_0 : \beta_p(t) = c_p, \forall t \in [0, 1]; \text{v.s. } H_1 : \beta_p(t) \neq c_p, \exists t \in [0, 1],$$

where $c_p$ is a constant of each $p$. If the line is covered, we accept that the coefficient is constant; otherwise, it is not. In Fig. 2, the curve indicates the expected solar intensity for two years, and illustrates an obvious seasonal pattern. The width of the $95\%$ SCB of $\beta_1(t)$ is so narrow that no horizontal line can be covered, and even a higher level of $98\%$ SCB cannot cover a horizontal line. We are confident that the solar power generation is time-varying, the same as the natural process.

As we center all the weather variables on their averages, the SCB of the $\beta_p(t)$ actually indicates the effect on the solar intensity. In each figure of Fig. 3 to 5, the zero line is not covered, while in Fig. 6 and 7, the zero line is covered by the $95\%$ SCB. Therefore, we can conclude that for a level of $95\%$, temperature, humidity and dew point have a strong effect on solar generation, but the effect from wind and precipitation are weak. Also, we accept $\beta_1(t)$ to $\beta_4(t)$ as time-varying coefficients, because a constant horizontal line cannot be covered entirely in those SCBs. Note the SCB associated with wind in Fig. 6 shows some variations. Although the zero line may not be covered by a narrower SCB, say $90\%$ SCB, at the $95\%$ significant level, we do not accept $\beta_5(t)$ as a non-zero function.The SCB associated with precipitation in Fig. 7 is also too wide for $\beta_6(t)$ to be accepted as a non-zero function.

TABLE II
RMS-ERRORS IN $watts/m^2$ FOR TLLE, SVM AND MLR

| | TLLE | SVM | MLR |
|---|---|---|---|
| RMS-Error | 22.59 | 32.71 | 53.35 |

### D. Comparisons with Other Models on Prediction Results

From the above discussions, we could exclude the variable of precipitation and simplify the model for better prediction. We use the model to predict the daily solar intensity for January and February in 2013. The weather information of the previous year is used as the weather forecast, and the time for prediction is set from day 366 to 423, using data around January and February in 2012. In other words, the predictions are made using the data around the same time in the previous year. The results are shown in Fig. 8. The upper graph is the prediction curve made by the time-varying local linear model (TLLM) and the actual observations; the lower one shows the results from SVM regression used in [6]. We also perform the multiple linear regression (MLR) as in (1). But the prediction is too poor to be shown as a comparison here.

From Fig. 8, we can see that the TLLE predicted curve tracks the actual observations better than the SVM regression. And it is also shown in Table II that the root mean squares error (RMS-Error) between the predicted series and the observations for TLLE is 22.59 $watts/m^2$, smaller than that of SVM and MLR, which are 32.71 $watts/m^2$ and 53.35 $watts/m^2$ respectively. Note that the SVM regression depending highly on the selection of the parameters and the kernels, and thus, is not practical in many cases lacking a comprehensive understanding of the real model. For example, the kernel function chosen for daily prediction is not guarantied to perform well in weekly prediction. However, the TLLE analyzes the model using simultaneous inference, which reflects the overall pattern of the dynamic pattern of the regression functions. Therefore, it can be used in many other applications, such as the hourly or weekly solar power generation forecast where the time scale is set in hour or week.

## V. CONCLUSIONS

In this paper, we proposed the simultaneous inference for weather dependent power generation from renewable energy sources, such as solar and wind. We first introduced the local linear model for time series, and presented the construction of the SCB for time-varying coefficients. We then performed the SCB analysis with a trace of solar intensity and weather

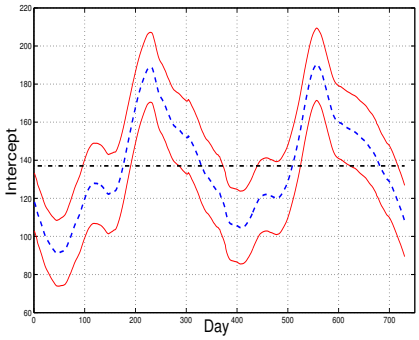


Fig. 2. 95% SCB for Intercept.

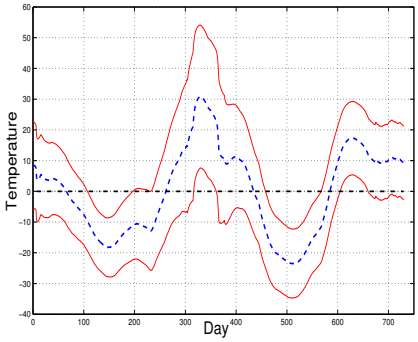


Fig. 3. 95% SCB for Temperature.

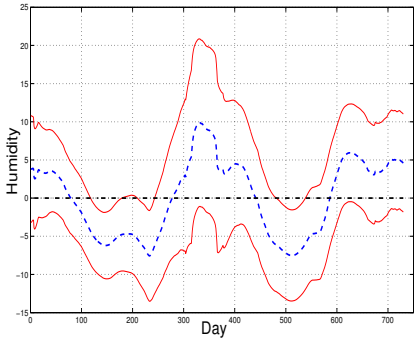


Fig. 4. 95% SCB for Humidity.

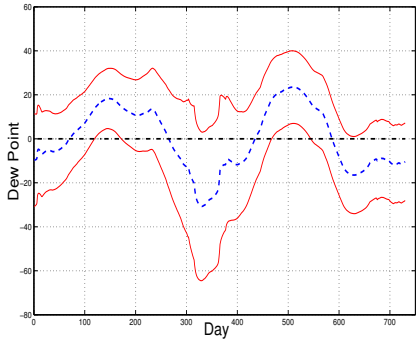


Fig. 5. 95% SCB for Dew Point.

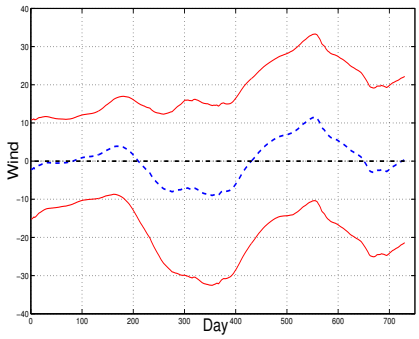


Fig. 6. 95% SCB for Wind.

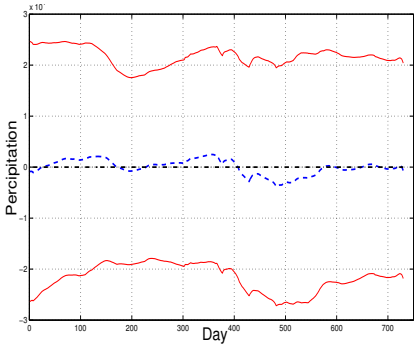


Fig. 7. 95% SCB for Precipitation.

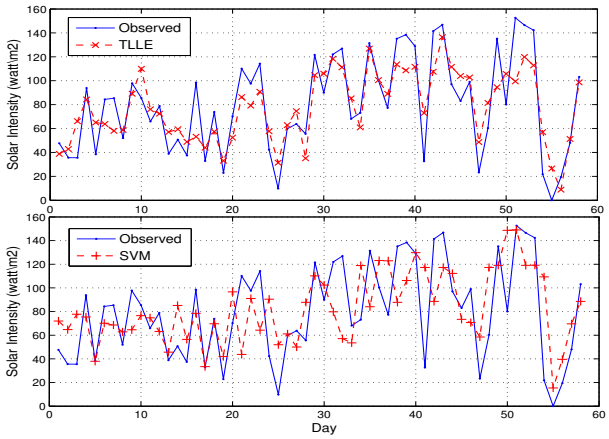


Fig. 8. Comparisons of predictions on solar intensity.

data to validate the efficacy of the proposed approach. The presented model was shown to outperform an existing method for solar intensity prediction.

## ACKNOWLEDGMENT

This work is supported in part by the US National Science Foundation under Grant CNS-0953513.

## REFERENCES

[1] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid – the new and improved power grid: A survey," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 944–980, Apr. 2012.

[2] Y. Wang, S. Mao, and R. Nelms, "Distributed online algorithm for optimal real-time energy distribution in the smart grid," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 70–80, Feb. 2014.

[3] M. Chaouch, "Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 411–419, Jan. 2014.

[4] V. Dordonnat, S.J. Koopman, and M. Ooms, "Dynamic factors in periodic time-varying regressions with an application to hourly electricity load modelling," *Comput. Stat. Data Anal.*, vol. 56, no. 11, pp. 3134–3152, 2012.

[5] H.S. Hippert, C.E. Pedreira, and R.C. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans. Power Syst.*, vol. 16, no. 1, pp. 44–55, Feb. 2001.

[6] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, "Predicting solar generation from weather forecasts using machine learning," in *Proc. IEEE SmartGridComm'11*, Oct. 2011, pp. 528–533.

[7] S. Zhu, M. Yang, M. Liu, and W.J. Lee, "One parametric approach for short-term jpdf forecast of wind generation," in *Proc. 2013 IEEE IAS Annual Meeting*, Orlando, FL, Oct. 2013, pp. 1–7.

[8] F. Ferraty and P. Vieu, *Nonparametric Functional Data Analysis: Theory and Practice.* New York: Springer-Verlag, 2006.

[9] W. Hardle, "Applied nonparametric regression," *Econometric Theory*, vol. 8, no. 3, pp. 413–419, Sept. 1992.

[10] G. Cao, L. Yang, and D. Todem, "Simultaneous inference for the mean function based on dense functional data," *J. Nonparametric Stat.*, vol. 24, no. 2, pp. 359–377, June 2012.

[11] Z. Zhou and W.B. Wu, "Local linear quantile estimation of nonstationary time series," *Ann. Statist.*, vol. 37, no. 5B, pp. 2696–2729, July 2009.

[12] D. Draghicescu, S. Guillas, and W.B. Wu, "Quantile curve estimation and visualization for nonstationary time series," *J. Comput. Graph. Statist.*, vol. 18, no. 1, pp. 1–20, 2009.

[13] Z. Zhou and W.B. Wu, "Simultaneous inference of linear models with time varying coefficients," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 513–531, Sept. 2010.

[14] G. Cao, "Simultaneous confidence bands for derivatives of dependent functional data," *Elec. J. Stat.*, vol. 8, no. 2, pp. 2639–2663, Dec. 2014.

[15] J. Fan and I. Gijbels, *Local Polynomial Modelling and Its Applications*. New York: Chapman and Hall, 1996.

[16] P. Craven and G. Wahba, "Smoothing noisy data with spline functions," *Numer. Math*, vol. 31, no. 4, pp. 377–403, Dec. 1979.

[17] E. Montgomery, C.and Peck and G. Vining, *Introduction to Linear Regression Analysis*, 4th ed. A John Wiley & Sons, Inc., July 1991.

[18] [online] Available: http://traces.cs.umass.edu/.