Abstract—Solar energy is an important source of renewable energy that can be harnessed using a range of evolving technologies such as solar heating, solar photovoltaic, solar thermal energy, solar architecture and artificial photosynthesis. The harnessed solar energy can be used in a wide range of applications like electricity production, fuel production, agriculture planning, water heating, transport, etc. The prediction is focusing in the Southern part of India and the solar light will be available from 8 to 9 months in a year in this region. So to utilize the solar energy in an efficient way the prediction is done. To predict the availability of solar energy a machine learning Temporal Gaussian Process Regression (TGPR) method has been used. It provides better result and also more robust when compared with the existing methods using ELM, SVM, etc. The predicted values could be used to measure and analyze the amount of energy that could be generated and in turn to identify the suitable solar based devices that can be installed in different locations.

Index Terms—Temporal Gaussian Process Regression (TGPR), machine learning, solar energy, solar devices.

I. INTRODUCTION

Solar energy is an important source of renewable energy which is currently under expansion in many countries. Solar energy production can be estimated from solar energy based systems and this estimation requires the prediction of available solar energy. The information on amount of energy present at a specific location during a specific time period can be derived from solar radiation data. This information can be used for designing and planning the solar energy based systems[1]. Due to the high cost and installation difficulties in measuring solar energy, these data are not always available. Therefore, there is a demand to develop alternative ways of predicting these data. Many machine learning techniques like Extreme learning machine (ELM), Support vector Regression (SVR), Artificial Neural Networks (ANN) are being used to predict the daily global solar radiation data[2]. Temporal Gaussian process regression is an alternative method to predict solar radiation data and it is proved to provide better results and also more accurate than the existing models. The solar parameters collected from the meteorological centre are used in solar radiation prediction[3]. The parameters used in the prediction process are maximum air temperature, relative humidity, minimum air temperature, atmospheric air pressure, wind speed, wind direction, etc.

Gaussian process regression is a method of interpolation for which the interpolated values are modeled by a Gaussian process governed by prior covariance. In this method, new data points can be constructed within the range of a discrete set of known data points. Under suitable assumptions on the priors, Gaussian process regression gives the best linear unbiased prediction of the values. Gaussian regression process can handle the limitation in the availability of training data. In addition to the good numerical performance and stability, it can adopt very flexible kernel functions, and it provides confidence intervals for the predictions[4].

II. RELATED WORKS

J. Zeng and W. Qiao[5] proposed a least-square (LS) support vector machine (SVM)-based model for short-term solar power prediction (SPP). The input of the model included historical data of atmospheric transmissivity in a novel two-dimensional (2D) form and other meteorological variables, including sky cover, relative humidity, and wind speed. The output of the model was the predicted atmospheric transmissivity, which then was converted to solar power according to the latitude of the site and the time of the day.

J. L. Chen et al[6] have presented the methods of monthly mean daily solar radiation estimation using support vector machines (SVMs) to examine the
feasibility of SVMs in estimating monthly solar radiation using air temperatures[7].

Tamer Khatib, et al[8] presented an assessment for the artificial neural network (ANN) based approach for hourly solar radiation prediction. The Four ANNs topologies were used including a generalized (GRNN), a feed-forward backpropagation (FFNN), a cascade-forward backpropagation (CFNN), and an Elman backpropagation (ELMNN). The three statistical values used to evaluate the efficacy of the neural networks were mean absolute percentage error (MAPE), mean bias error (MBE) and root mean square error(RMSE). Prediction results show that the GRNN exceeds the other proposed methods.

J. Verrelst, et al[9] proposed a model for Retrieval of Vegetation Biophysical Parameters Using Gaussian Process Techniques. This model evaluated state-of-the-art parametric and nonparametric approaches for the estimation of leaf chlorophyll content (Chl), leaf area index, and fractional vegetation cover from space. The proposed GP method was able to build robust relationships between the parameter of interest and only a few bands.

L. Pasolli, F. Melgani, E. Blanzieri[10] have explored the effectiveness of a novel regression method in the context of the estimation of biophysical parameters from remotely sensed imagery as an alternative to state-of-the-art regression methods like those based on artificial neural networks and support vector machines. Moreover, it handled the problem of limited availability of training samples, typically encountered in biophysical parameter estimation applications.

M. Hassan, A. Bernak[11] have developed the model by integrating wireless sensing nodes with ambient energy harvesting capability to overcome limited battery power budget constraint and extending effective operational time of sensor network. An efficient algorithm for solar energy prediction based on additive decomposition (SEPAD) model was used. In this model, both seasonal and daily trends along with Sun’s diurnal cycle are individually considered.

Jose Luis Guinon, et al[12] presented the use of Mathcad software for the implementation and analysis of the moving average and Savitzky-Golay filters. By means of the Mathcad software, moving average and Savitzky-Golay filters were successfully applied to the smoothing of photochemical and electrochemical reactor data.

III. SOLAR RADIATION PREDICTION

The Fig 1 depicts the steps involved in daily global solar irradiation prediction system. It consists of five stages which includes preprocessing, learning, prediction and prediction analysis. The solar sensors collect the solar irradiation data. The collected parameters include air temperature, relative humidity, air pressure, wind speed, wind direction, global horizontal irradiance, diffuse normal irradiance, etc. The collected data set contain around 54 parameters of information out of which only a few parameter are required for daily global solar radiation prediction.

Also, the use of noisy data will lead to failure in the prediction process. In preprocessing, filtering technique is applied to remove the noise from the collected data values.

The relationship and dependency between the solar variables is learnt by finding the covariance between two variables. Temporal Gaussian process regression method is used to predict the future data values using the given set of known data values based on time series.

The predicted solar irradiation values are correlated with actual or observed solar irradiation values for accuracy based on root mean square error, mean absolute error and mean bias error. The amount of energy that could be generated using different sizes of photovoltaic cell can be measured and analyzed to identify the solar energy based devices with request to the power requirement.

A. Preprocessing

The global solar data around the Chennai region is collected for prediction process. The collected parameters
like minimum air temperature, relative humidity, air pressure from the solar sensor obtained for five years is given as input to the preprocessing stage. The collected data set may contain too much of information out of which only a few data are required for daily global solar radiation prediction. Also, the use of noisy data will lead to failure in the prediction process. In preprocessing, moving average filtering technique is applied to remove the noise from the collected data values[13].

Moving average filter

This computation involves the analysis of data points by creating a series of averages of different subsets of the full data set. The moving average approach is commonly used with time series data to smooth out the fluctuations in the data[12].

\[ y = \text{filter}(b, a, x) \]  
\( (1) \)

creates filtered data \( y \) by processing the data in vector \( x \) with the filter described by vectors \( a \) and \( b \). The filter function is a general tapped delay-line filter, described by the difference equation

\[ a(1)y(n)=b(1)x(n)+b(2)x(n-1)+…+b(N_b)x(n-N_b+1) \]
\[ -a(2)y(n-1)-…-a(N_a)y(n-N_a+1) \]

\( (2) \)

Here, \( n \) is the index of the current sample, \( N_a \) is the order of the polynomial described by vector \( a \), and \( N_b \) is the order of the polynomial described by vector \( b \). The output \( y(n) \) is a linear combination of current and previous inputs, \( x(n)x(n-1)… \), and previous outputs, \( y(n-1) y(n-2)… \).

The Fig. 2 shows the erroneous data of maximum air temperature which is analyzed and filtered as shown in Fig. 3.

B. Learning correlation

Data can be analyzed to learn how each parameters affect the other and the solar intensity. It was found that the parameters like sky cover, relative humidity are highly correlated with each other and with solar intensity. Also the parameters like temperature, dew point and wind speed are partially correlated with each other and with solar intensity[14].

The relationship and dependency between the parameters are learnt by finding the covariance between the parameters. Based on the learnt knowledge, the parameters are chosen to use in the prediction model.

Correlation is a scaled version of covariance, the two parameters always have the same sign (positive, negative, or 0). When the sign is positive, the variables are said to be positively correlated; when the sign is negative, the variables are said to be negatively correlated; and when the sign is 0, the variables are said to be uncorrelated.

Covariance is a measure of how much two random variables change together. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, i.e., the variables tend to show similar behavior, the covariance is positive.
In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, i.e., the variables tend to show opposite behavior, the covariance is negative. The sign of the covariance therefore shows the tendency in the linear relationship between the variables. The covariance and correlation measure a certain kind of dependence between the variables. The sample covariance of $N$ observations of $K$ variables is the $K$-by-$K$ matrix $\mathbf{Q} = [q_{jk}]$ with the entries

$$q_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

(3)

which is an estimate of the covariance between variable $j$ and variable $k$.

The sample mean and the sample covariance matrix are unbiased estimates of the mean and the covariance matrix of the random vector $\mathbf{X}$, a row vector whose $j$th element ($j = 1, \ldots, K$) is one of the random variables. The reason the sample covariance matrix has $N-1$ in the denominator rather than $N$ is that the population mean $E(\mathbf{X})$ is not known and is replaced by the sample mean $\bar{X}$.

The correlation between the parameters can be calculated using the formula,

$$\text{kernel\_function} = \sigma_f^2 \exp((-2)^p(x-x2)/(-2)^p2))$$

(4)

Where $x$ and $x2$ are the input parameters like air temperature and relative humidity and kernel function computes the covariance between parameters.

C. Prediction

Solar energy prediction is done using Gaussian process regression model. It is a method of interpolation for which the interpolated values are modeled by a Gaussian process governed by prior covariance. In this method, new data points can be constructed within the range of a discrete set of known data points. Under suitable assumptions on the priors, Gaussian process regression gives the best linear unbiased prediction of the values[15]. The main idea is to predict the value of the function at a given point by computing a weighted average of the known values of the function in the neighborhood of the point.

It is assumed that for a Gaussian process $f$ observed at coordinates $x$, the vector of values $f(x)$ is just one sample from a multivariate Gaussian distribution of dimension equal to number of observed coordinates $|x|$. Therefore under the assumption of a zero-meaned distribution,

$$f(x) \sim N(0, K(0,x,x'))$$

(5)

where $K(0,x,x')$ is the covariance matrix between all possible pairs $(x,x')$ for a given set of hyperparameters $\theta$. Maximizing this marginal likelihood towards $\theta$ provides the complete specification of the Gaussian process $f$.

Having specified $\theta$ making predictions about unobserved values $f(x^*)$ at coordinates $x^*$ is then only a matter of drawing samples from the predictive distribution

$$p(y^*|x^*,f(x),x) = N(y^*|A,B)$$

(6)

where $K(0,x^*,x)$ is the covariance between the new coordinate of estimation $x^*$ and all other observed coordinates $x$ for a given hyperparameter vector $\theta$, $K(0,x,x')$ and $f(x)$ are defined as before and $K(0,x^*,x^*)$ is the variance at point $x^*$ as dictated by $\theta$. It is important to note that practically the posterior mean estimate $f(x^*)$ is just a linear combination of the observations $f(x)$; in a similar manner the variance of $f(x^*)$ is actually independent of the observations $f(x)$. A known bottleneck in Gaussian process prediction is that the computational complexity of prediction is cubic in the number of points $|x|$ and as such can become unfeasible for larger data sets. Works on sparse Gaussian processes, that usually are based on the idea of building a representative set for the given process $f$.

![Fig. 4. Monthly Average Solar Radiation Prediction](image)

The Fig. 4 shows the monthly average solar radiation prediction for one year with an error of 4.5998%.

D. Prediction Analysis

The predicted solar irradiation values are compared with actual or observed solar irradiation values for accuracy based on root mean square error or root mean square prediction error.
square deviation (RMSD), mean absolute error (MAE) and mean bias error (MBE) [14].

**RMSD:**

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between predicted values and actually observed values.

$$\text{RMSD} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$  \hspace{1cm} (7)

The RMSD of predicted values \(\hat{y}_i\) for times \(i\) of a regression’s dependent variable \(y\) is computed for \(n\) different predictions as the square root of the mean of the squares of the deviations.

**MAE:**

The mean absolute error (MAE) is a quantity used to measure how close predictions are to the eventual outcomes. The mean absolute error is given by

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$ \hspace{1cm} (8)

The mean absolute error is an average of the absolute errors \(e_i = |\hat{y}_i - y_i|\), where \(\hat{y}_i\) is the prediction and \(y_i\) the true value.

**TABLE I**

<table>
<thead>
<tr>
<th>METHOD</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELM</td>
<td>4.22</td>
<td>2.11</td>
</tr>
<tr>
<td>SVR</td>
<td>3.23</td>
<td>2.12</td>
</tr>
<tr>
<td>GPR</td>
<td>2.15</td>
<td>1.27</td>
</tr>
<tr>
<td>TGPR</td>
<td>0.865</td>
<td>0.724</td>
</tr>
</tbody>
</table>

REGRESSION MODELS

IV. CONCLUSION

Solar energy prediction provides support for measuring and analyzing solar energy. Gaussian process regression based on time series is used in the daily global solar energy prediction. This method is proved to provide better results and also more robust when compared with the existing models. The average RMSD and MAE for the comparison between observed and predicted solar irradiation are 0.865 and 0.724, respectively. The predicted values could be used to measure and analyze the amount of energy that could be generated and in turn to identify the suitable solar based devices that can be installed in different locations.

REFERENCES

[3]. Bjorn Wol, Elke Lorenz, Oliver Kramer "Statistical Learning for Short-Term Photovoltaic Power Predictions".
U. Divya received the B.Tech degree in Information Technology in 2013 and currently doing her M.E degree in Computer Science and Engineering. She has Published a paper titled on “A STUDY ON BIG DATA” in International Journal of Engineering and Research and Technology and also Published a paper titled on “SURVEY ON MACHINE LEARNING APPROACHES FOR SOLAR IRRADIATION PREDICTION” in International Journal of Engineering Sciences and Research Technology.

Dr. P. Chitra has been working as Professor in Department of Computer Science and Engineering, Prathyusha Institute of Technology and Management, Tamilnadu, India. She has completed doctorate from Anna University, Chennai in 2013. She has more than 15 years of experience as academician and around six years of experience in industry. She has organized various IEEE/ISTE sponsored workshops, seminars and conferences. She is a life member in ISTE, IACSIT and IAENG. Her area of research includes Data Mining, Image Processing, Information Security, Big Data Analytics, etc.