Calculation of Solar Radiation by Using Regression Methods

Ö Kızıltan and M Şahin

Electrical and Electronic Engineering, Engineering Faculty, Siirt University, 56100, Siirt, Turkey

Abstract. In this study, solar radiation was estimated at 53 location over Turkey with varying climatic conditions using the Linear, Ridge, Lasso, Smoother, Partial least, KNN and Gaussian process regression methods. The data of 2002 and 2003 years were used to obtain regression coefficients of relevant methods. The coefficients were obtained based on the input parameters. Input parameters were month, altitude, latitude, longitude and land-surface temperature (LST).The values for LST were obtained from the data of the National Oceanic and Atmospheric Administration Advanced Very High Resolution Radiometer (NOAA-AVHRR) satellite. Solar radiation was calculated using obtained coefficients in regression methods for 2004 year. The results were compared statistically. The most successful method was Gaussian process regression method. The most unsuccessful method was lasso regression method. While means bias error (MBE) value of Gaussian process regression method was 0,274 MJ/m^2, root mean square error (RMSE) value of method was calculated as 2,260 MJ/m^2. The correlation coefficient of related method was calculated as 0,941. Statistical results are consistent with the literature. Used the Gaussian process regression method is recommended for other studies.

1. Introduction

Solar radiation (SR) predictions are performed in a geographical region, operating conditions of many solar energy systems, which are in design and development stage, can be simulated in the related region [1]. More clearly, long-term SR predictions are an essential parameter for engineering applications such as modelling solar power plants, modelling photovoltaic cells, and modelling solar heating systems [2]. For this reason, it is necessary to accurately prediction SR, which is excessive valuable information. In the previous years, SR prediction models have been developed in harmony with parameters such as air temperature, humidity, sunshine duration, and cloud coverage, which were measured from conventional meteorological stations and which were evaluated indirectly as a function of SR [3]. These models can be categorized into two groups, namely parametric methods such as Angstrom [4] and nonparametric methods that are based on artificial intelligence [5, 6]. In the literature, it has been observed that SR information in a specific location can be estimated using these models. However, it may not be possible to obtain accurate and continuous data from every station because maintenance and calibration of measuring devices, which are installed in meteorological stations for SR measurement, are difficult and installation costs are high [7]. In addition, the fact that the number of meteorological stations are limited especially in developing countries and inefficient recording of data due to device malfunctions constitute another limitation in obtaining SR data [8]. Thus, these limitations forced researchers toward a tendency to develop alternative estimation methods and find more reliable data sources for the regions where SR data cannot be directly measured or stations are insufficient. In recent years, satellite-based remote sensing (RS) techniques are widely used as an alternative method and as a data source for SR estimations [7, 9]. One of the most especial advantages of RS is that it is a reliable and rapid

Published under licence by IOP Publishing Ltd

method for obtaining up-to-date and continuous information about wide geographical areas. In addition to this, satellite-based RS provides opportunity to perform SR estimations in rural, mountainous, and remote places where meteorological stations are insufficient. Calculations of solar radiation can be made by with various methods. It has been made with regression models using data from ground stations. Regression analysis, to explain the relationship between variables as a functional and this relationship is described by a model. It used to measure the magnitude of the relationship between variables. And it enables us to find the cause-effect relationships between variables. To sum up; Regression analysis is to find the proper function of the data table.

2. The Machine Learning

The learning machine, which was developed by Huang et al. [10], is a novel learning algorithm for single hidden layer feed forward networks (SLFNs). It has been widely used for the solution of estimation problems in many different fields [11, 12]. There are some advantages of the ELM algorithm. (i) It is easy to use, and no parameters need to be tuned except predefined network architecture and thus avoid many difficulties faced by gradient- based algorithms such as learning rate, learning epochs, and local minima. (ii) It is proven to be a faster learning algorithm compared with other conventional learning algorithms such as back propagation (BP) algorithm. Most training can be accomplished in seconds and minutes (for large-scale complex applications), which might not be easily obtained using other traditional learning methods. (iii) It possesses similar high generalization performance as BP [13] whatever the application area, the method of analyzing a lot of data available in machine learning methods to make predictions about the future

3. Methodology

In this study, Solar Radiation was predicted for 53 locations by using Linear, Ridge, Lasso, Smoother, Partial least, KNN and Gaussian process regression methods. These methods were used of data 2002 and 2003 years and accuracy of the mentioned methods was tested yearly. The data set of each year was consisting of 12 months. Whereas the values of month, altitude, latitude, longitude and land surface temperature were used as input for developing models, SR was obtained as output. Land surface temperature (LST) was taken from radiometry of NOAA-AVHRR satellite. Estimated solar radiation data were compared with actual data that were obtained from satellite data.

Learning Machine has a very valuable method of modelling tools, many of them with a proven track record in applications. K-NN is a non-parametric method used for categorization and regression. [14].LR measures the relationship between the categorical dependent variable and one or more independent variables by prediction probabilities using a logistic function, which is the cumulative logistic distribution.

3.1. Linear regression

LR is a machine learning method which finds the relation between the dependent variable and the independent variable or variables [15]. This methods depends on the assumption of the relationship of this variables are polynomial as in Eq. 1.

$$y_i = \beta_1 x_{i_1} + \cdots + \beta_p x_{i_p} + \varepsilon_i \tag{1}$$

Where, equation 1 can also be shown as a format of equation 2.

$$y = X\beta + \varepsilon \tag{2}$$

Where y is dependent variable, X is independent variable, β is parameter vector and ε is error term shown below.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}, X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix} \qquad (3)$$

3.2. Ridge Regression

For the ease of presentation, we rewrite regression Eq. (1) into the following matrix form:

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{e} \qquad (4)$$

where **y** is a (n×1) vector of the observation values of the dependent variable, X is a (n×p) matrix of the p explanatory variables or p−1 explanatory variables with the first column being a column vector of 1s as the regression intercept, ˇ is a (p×1) vector of the unknown regression coefficients, and **e** is a (n×1) vector of normally distributed errors with E(e) = 0 and Var (e) σ^2In where In is a (n×1) vector with every element equals to (1).

The ordinary least square (OLS) regression is the fundamental method in the regression family. Though OLS regression is one of the oldest methods for cost estimation, it is still being continuously improved or frequently applied by many researchers. [16]

$$\hat{\beta} = (X'X)^{-1} X'y \qquad (5)$$

The key properties of $\hat{\beta}$ are (7) it is unbiased, which is E ($\hat{\beta}$) = β and it has the minimum variance among all linear unbiased estimators, which is assured by Gauss–Markov theorem

Mathematically, ridge regression parameter estimation has the following form:

$$\hat{\beta}(\mathrm{k}) = (X'X + \mathrm{kIn})^{-1} X' \mathrm{y}, \ \mathrm{k} > 0 \qquad (6)$$

$\hat{\beta}$ (k) in (9) is obtained by minimizing the combination of both the sum of square error and the norm of the $\hat{\beta}$ vector

$$\min(y - \hat{\beta})'(y - x\hat{\beta}) + k\hat{\beta}'\hat{\beta} \qquad (7)$$

The norm $\hat{\beta}'\hat{\beta}$ is the length of the vector $\hat{\beta}$ in a vector space. The purpose of adding one norm term to the sum of square errors is to restrict the values of $\hat{\beta}$ into a certain boundary. It is noted that the methodology of minimizing.

3.3. Lasso Regression

In statistics and statistical and machine learning, lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. It was introduced by Robert Tibshirani in 1996 based on Leo Breiman’s Nonnegative Garrote.[16][17] lasso was originally formulated for least squares models and this simple case reveals a substantial amount about the behavior of the estimator, including its relationship to ridge regression and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding.

Lasso regularization can be extended to a wide variety of objective functions such as those for generalized linear models, generalized estimating equations, proportional hazards models, and M-estimators in general, in the obvious way. [16] Given the objective function

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i, y_i, \alpha, \beta) \tag{8}$$

the lasso regularized version of the estimator will be the solution to

$$\min_{\alpha,\beta} \frac{1}{N}^{2} \sum_{i=1}^{N} f(x_{i,}y_{i,}\alpha,\beta) \text{ subject to } \lVert\beta\rVert_1 \le t \tag{9}$$

Where only is penalized while is free to take any allowed value, just as was not penalized in the basic case.

3.4. Kernel Smoother regression

Kernel-based methods are very popular in non-parametric estimation methods. The goal is, to find a non-linear functional relationship between random variables and [18]. Suppose that and are random variables defined in, the non-parametric regression model in equation 7.

$$Y(X) = \Psi(X) + \varepsilon \tag{10}$$

Nadaraya - Watson kernel method smoothed regression function provides estimation based on the weighted average of local values of in equation (8)

$$\Psi^{n}(X) = \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)} \tag{11}$$

Where is a kernel, usually chosen from a positive value of the symmetric functions such as Epanechnikov, Tri-cube and Gaussian kernels and is zero outside of defined inferable area [19]. h is the bandwidth and also scaling factor parameter and controlling the smoothness of regression function $\Psi(X)$.

3.5 Partial least squares regression

PLS is a bilinear calibration method which compresses by reducing a large number of measured collinear spectral data to a few non-correlated principal components (PCs) [20]. The aim of linear PLC is to reduce the data a few latent variables, such as t_j and u_j ($j = 1,\dots, A$; where is the number of hidden variables) and then to develop a regression model to reveal the relationship between t_j and u_j [21] is in equation 4

$$u_j = b_j t_j + e_j \tag{12}$$

where e_j is the error vector, and b_j is represented by the $\hat{b}_j = \left(t_j^T t_j\right)^{-1} t_j^T u_j$ and Latent variables are obtained by $t_j = X_j w_j$ and $u_j = Y_j q_j$, where w_j and q_j of unit length and t_j and u_j must have a maximum of the covariance. X_j is defined

$$X_{j+1} = X_j - t_j p_j^T \tag{13}$$

Where $X_1 = X$ and $p_j = X_j^T t_j / \left(t_j^T t_j\right)$ and $Y_{j+1} = Y_j - b_j t_j q_j^T$ where $Y_1 = Y$.

If $\hat{u}_j = \hat{b}_j t_j$ then matrices X and Y can be written ;

$$X = \sum_{j=1}^{A} t_j p_j^T + E \quad \text{and} \quad Y = \sum_{j=1}^{A} \hat{u}_j q_j^T + F \tag{14}$$

where E and F are residuals of X and Y

3.6 K Nearest Neighbors Regression

k-NN, which converges the output to the mean of k nearest neighbors of input at feature space, is in equation 3**[**22**].**

$$f_{knn}(x) = \frac{1}{K} \sum_{i \in N_{K(x)}} y_i \tag{15}$$

where; $f : x \rightarrow y$ is regression function, x is input, y_i is known output values and $N_{K(x)}$ is the indeces of k nearest neighbor of x .

3.7 Gaussian Process regression

GPR is a machine learning process which depends on Bayesian probability model. This model, which is shown in equation 9, assumes that the random variables are normal distributed and therefore is based on mean and covariance matrix of random variable for regression [23].

$$E\{y \backslash L, x\} = \int y p(f \backslash L, x) dy \tag{16}$$

where; $E\{y \backslash L, x\}$ is expectation of output conditioned to L, x; y is best estimation output, x is input, L is quantity condition, $p(f \backslash L, x)$ is predictive distribution and is assumed that predictive distribution is normal (Gaussian) distribution.

So;

$$p(f \backslash L, x) \sim N(\mu_x, \sigma_x^2) \tag{17}$$

where; $\mu_x = k_x^T [K + \sigma_n^2 I]^{-1} . y + b$ and $\sigma_x^2 = k(x, x') - k_x^T [K + \sigma_n^2 I]^{-1} k_x$ (18)

$k(x, x')$ is covariance matrix of training samples, K is covariance vector between the training samples and k_x is covariance vector of sample$x$. b is bias factor, σ_n is the noise variance and I is identity matrix.

4. Results

In this study, SR was calculated as related to 53 locations. Regression methods were used to make the calculations. Linear, Ridge, Lasso, Smoother, Partial least, KNN and Gaussian process regression methods were used to calculate SR. Locations of the 2002 and 2003 values were used to obtain the regression coefficients. Input parameters were month, altitude, latitude, longitude and land-surface temperature (LST).The values for LST were obtained from the data of the National Oceanic and Atmospheric Administration Advanced Very High Resolution Radiometer (NOAA-AVHRR) satellite. After obtaining the coefficients of the regression methods were used for 2004 of locations. The obtained SR values are evaluated with actual values statistically. R, MBE and RMSE methods were used statistical methods. These methods depending on the values are shown in Table1.

Table 1. Regression of Methods MBE, RMSE and R values.

METHOD	**MBE(MJ/m^2)**	**RMSE(MJ/m^2)**	**R**
Linear Regression	-2,732	5,742	0,660
Lasso Regression	-1,970	5,532	0,601
K.SmootherRegression	0,952	6,079	0,711
Partial Least squares reg.	-2,079	5,679	0,639
k-NN Regression	-0,998	3,500	0,864
Gaussian Process Regression	0,274	2, 260	0,941
Ridge Regression	-3,358	5,908	0,648

When table 1 examined, the smallest correlation coefficient value have been found 0,601 MJ/m^2.This result have been obtained by using Lasso Regression. The biggest correlation coefficient have been evaluated 0,901 MJ/m^2 . The obtained result have been got using Gaussian Process regression. On the other side, correlation coefficient values of other methods is changeable values between 0,639 MJ/m^2 and 0,864 MJ/m^2 .The best MBE value have been calculated 0,274 MJ/m^2.This result have been got by Gaussian Process regression methods. The worst MBE value is 358 MJ/m^2. Related Value is provided from the ridge regression method. Other statistical methods are used the RMSE method. Accordingly largest RMSE value of 6.079 MJ / m^2 is calculated. This value is obtained from the kernel regression smoothing method. This result is very important because the highest error of solar radiation calculations have been obtained by smooth Kernel regression method. The smallest RMSE value is 2,260 MJ/m^2, which have been got by using Gaussian Process regression method. So, it proved that this is the most successful method. This results shows that, Linear, Ridge, Lasso, Smoother, Partial least, KNN and Gaussian process regression methods are enough for correctly calculation. Depending on the location of this method, MBE, RMSE and R values are shown in table 2

Table 2. Statistical results of locations

LOCATION	**MBE(MJ/m^2)**	**RMSE(MJ/m^2)**	**R**
Amasya	-0,600	1,411	0,984
Antakya	-1,091	1,372	0,990
Antalya	-1,916	2,244	0,986
B.Kesir-Gönen	-0,555	1,857	0,960
Bingöl	2,631	3,541	0,939
Bitlis	-1,393	2,392	0,953
Denizli	-1,655	2,476	0,980
Edirne	-0,094	2,787	0,959
G.antep	-1,201	2,081	0,967
Gümüşhane	-0,812	1,344	0,988
Hakkari	2,421	4,220	0,803
Iğdır	0,270	2,418	0,917
İst-Göztepe	1,190	2,179	0,962
			0,962
İzmir	-0,339	2,034	
Konya	1,139	1,824	0,974
Malatya	0,165	1,182	0,988
Mersin	0,296	0,972	0,989
Muş	0,722	2,463	0,944
Niğde	0,943	1,651	0,978
Ordu	0,959	2,693	0,891

Rize	1,493	2,417	0,904
Siirt	-1,256	1,960	0,967
Silifke	0,068	0,693	0,994
Sinop	3,520	4,358	0,941
Ş.Urfa	2,602	3,163	0,963
Yalova	0,480	2,369	0,941
Yozgat	-0,485	1,873	0,977
Zonguldak	0,804	1,652	0,979
Adana	-1,125	1,930	0,966
Adıyaman	-1,215	2,097	0,977
Ağrı	-0,880	1,922	0,963
Aksaray	1,510	2,033	0,988
Akşehir	-0,198	1,533	0,974
Ankara	0,020	1,082	0,986
Artvin	-0,331	1,885	0,947
Aydın	-1,910	2,430	0,980
Batman	1,279	2,637	0,950
Bilecik	0,584	1,546	0,983
Birecik	1,442	1,874	0,983
Burdur	2,278	2,452	0,991
Çanakkale	0,253	1,878	0,958
Çorum	0,358	1,719	0,979
Elazıg-Bölge	-0,947	1,947	0,973
Erzincan	0,052	1,845	0,987
Isparta	2,049	2,964	0,947
K.maras	0,326	1,473	0,977
Karaman	1,506	1,884	0,987
Karataş	-0,613	1,452	0,972
Kars	-0,607	1,406	0,985
Kastamonu	-0,874	2,676	0,965
Kayseri	-0,708	1,549	0,988

When table 2 is investigated, the lowest correlation coefficient value have been evaluated 0,803 MJ/m^2. This result has been provided from Hakkari location. The highest correlation coefficient value has been found 0,994 MJ/m^2. The related value has been obtained from Silifke location. In this study, MBE values also calculated. The best MBE value is is 0,020 MJ/m^2 which have been got from Ankara location. The worst MBE value has been found 3,520 MJ/m^2. This value has been provided from Sinop location. The maximum value for the calculation of RMSE is calculated 4,358 MJ / m^2. This calculation is performed at the location of Sinop. It proved that Sinop location is the most unsuccessful according to other calculations of SR. The smallest RMSE value has been calculated to as 0,693 MJ / m^2 .The related value is belonging to Silifke location. According to Gaussian Process regression method, most successful calculation of SR was provided from Silifke location. SR variation of Silifke location monthly was shown in Figure 1

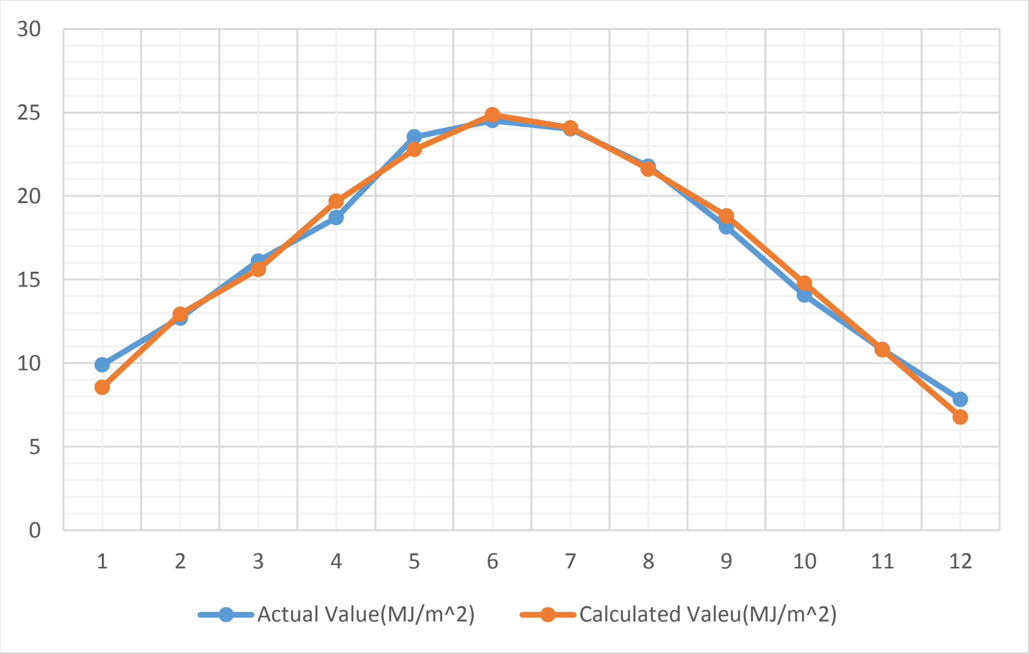

Figure 1.GSR change of Silifke location monthly

When we examined the GSR variation of location, we can observed that actual value and calculated value is very close to each other. The second, sixth, seventh, eight and eleventh month's value are more conformable that it has been observed. We cannot say the same for the other months.

5. Conclusion

In this study, calculation of SR was the different used regression methods. These methods are Linear, Ridge, Lasso, Smoother, Partial least, KNN and Gaussian process regression methods. It used to calculate SR. Gaussian process regression method has been the most successful method calculation of SR. On the other hand Linear Regression method was found to be the most unsuccessful method. The other methods gave agreeable results. If researcher studies calculation of SR we may suggest them to use Gaussian process regression method

References

[1] Marti P Gasque M, Improvement of temperature based solar radiation estimation through exogenous data assistance. *Energy Conversion and Management* 2011; **52** pp: 990– 1003.

[2] Koca A Oztop HF Varol Y Koca GO, Estimation of solar radiation using artificial neural networks with different input parameters for Mediterranean region of Anatolia in Turkey. Expert Systems with Applications 2011; **38** pp: 8756–8762.

[3] Ehnberg JSG, Bollen MHJ. Simulation of global solar radiation based on cloud observations. Solar Energy 2005; **78** pp: 157–62.

[4] Akinoglu BG, Ecevit A. Construction of a quadratic model using modified Angstrom cofficients to estimate global solar radiation. Solar Energy 1990; **45** pp: 85–92

[5] Sözen A, Arcaklıoğlu E, Özalp M, Kanıt EG. Use of artificial neural networks for mapping of solar potential in Turkey. Applied Energy 2004; 77 pp: 273-86

[6] Mellit A. Artificial Intelligence technique for modelling and forecasting of solar radiation data: a review. International Journal Artificial Intelligence and Soft Computing 2008; **1** pp: 52– 76.

[7] Qin J, Chen Z, Yang K, Liang S, Tang W. Estimation of monthly-mean daily global solar radiation based on MODIS and TRMM products. Applied Energy 2011; **88** pp: 2480–2489

[8] Chen JL, Liu HB, Wu W, Xie DT. Estimation of monthly solar radiation from measured temperatures using support vector machines: a case study. Renewable Energy 2011; **36** pp: 413–420.

[9] Kandırmaz HM, Yeğingil I, Peştemalcı V, Emrahoğlu N. Daily global solar radiation mapping of Turkey using Meteosat satellite data. International Journal of Remote Sensing 2004; **25** pp: 2159–2168.

[10] Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. Neurocomputing 2006; **70** pp: 489–501...

[11] Malathi V, Marimuthu NS, Baskar S, Ramar K. Application of extreme learning machine for series compensated transmission line protection. Engineering Application of Artificial Intelligence 2011; **24** pp: 880–887

[12] Yang H Zhao J Wang D and Dong Z, Predicting the probability of ice storm damages to electricity transmission facilities based on ELM and Copula function. Neurocomputing 2011; **74** pp: 2573–2581.

[13] Suresh S Saraswathi S Sundararajan N, Performance enhancement of extreme learningmachine for multi category sparse data classification problems. Engineering Application of artificial Intelligence 2010; **23** pp: 1149–1157.

[14] Altman N S, An introduction to kernel and nearest-neighbour nonparametric regression The American Statistician 1992; **46 (3)** pp: 175–185.

[15] Almorox J Hontoria C, Global Solar Radiation estimating using sunshine duration in Spain 2004; **45** pp: 1529-1535

[16] Tibshirani R, Regression Shrinkage and Selection via the lasso. Journal of the Royal Statistical Society. Series B (methodological) **58 (1)** pp Wiley: 267 88.

[17] Breiman, L, 1995 Better Subset Regression Using the Nonnegative Garrote. Techno metrics **37 (4)** pp. Taylor & Francis, Ltd.: 373–84. Doi: 10.2307/1269730.

[18] Bacher P Madsen H Aalborg H, Online short-term solar power forecasting 2009; **83** pp: 1772-1783.

[19] Şenkal Ö, Modeling of Solar Radiation Using Remote sensing and artificial neural network in Turkey December 2010; **35** pp: 4795-4801.

[20] Seshu D V Cady F B, Response of Rice to Solar Radiation and Estimated from International Yield Trials 1983; **40** pp: 649-654

[21] Guillermo P P Liliana N Carlos A V Maria A S, Estimating daily Solar Radiation in the Argentine Pampas 2004; **123** pp: 41-53.

[22] Paoli C Voyant C Muselli M Nivet M L, Forecasting of preprocess daily solar radiation time series using neural networks. December 2010; **84** pp.2146-2160

[23] Dinçer İ Dilmac Ş Ture E Edin M, A simple technique for estimating solar radiation parameters and its application for Gebze 1996; **37** pp: 183-198