Hittite Journal of Science and Engineering, 2020, 7 (1) 35–40
ISSN NUMBER: 2148–4171
DOI: 10.17350/HJSE19030000169

Solar Power Prediction with an Hour-based Ensemble Machine Learning Method

Seyda Ertekin
Middle East Technical University, Department of Computer Engineering, Ankara, Turkey

ABSTRACT

In recent years, the share of solar power in total energy production has gained a rapid increase. Therefore, prediction of solar power production has become increasingly important for energy regulations. In this study we proposed an ensemble method which gives competitive prediction performance for solar power production. This method firstly decomposes the nonlinear power production data into components with a multi-scale decomposition technique such as Empirical Mode Decomposition (EMD). These components are then enriched with the explanatory exogenous feature set. Finally, each component is separately modeled by nonlinear machine learning methods and their results are aggregated as final prediction. We use two different training approaches such as Hour-based and Day-based, for predicting the power production at each hour in a day. Experimental results show that our ensemble method with Hour-based approach outperform the examined machine learning methods.

Article History:
Received: 2019/11/23
Accepted: 2020/03/17
Online: 2020/03/26

***Correspondence to**: Şeyda Ertekin,*
Department of Computer Engineering,
Middle East Technical University,
06680, Ankara, Turkey
E-mail: sertekin@metu.edu.tr,
Phone: +90 312 210 5509,
Fax: +90 312 210 5544.

Keywords:
Solar power, Time series forecasting, Machine learning, Ensemble methods, Empirical mode decomposition

INTRODUCTION

Energy sustainability is crucial for economic growth and financial development. Therefore, supplying energy through sustainable resources plays an important role in national economic strategies. It is well known that renewable energy sources such as solar power are sustainable in long-term timeline as compared to fossil resource [1]. With the developing technology, the cost of solar panels such as Photo-Voltaic, which directly produce electricity from solar energy, is decreasing. Thus, in developing countries such as Turkey, solar energy production comes to the increasingly important level.

The total installed capacity of solar power plants in Turkey increased by 57% over the last two years and has reached to 5 GW [2]. With the planned capacity increase, the installed capacity is expected to exceed 30 GW in the next 10 years [3]. In short, the share of total production of solar energy production in Turkey is increasing rapidly. Therefore, forecasting solar power production for making decision in power system scheduling and grid regulation plays an important role in Turkey's electric industry.

Solar power forecasting models are categorized according to model designs and inputs used by models. The Numerical Weather Forecast (NWP) model has been widely applied for solar radiation estimation [4,5]. This approach uses mathematical models of weather conditions for prediction. This method is often used for short-term (such as 6-hour) forecasts and model performance varies depending on weather data such as cloud cover and temperature [6,7].

Time series models aim to predict future data from past observed data using statistical and machine learning techniques. In literature, there are various time series methods [8, 9, 10, 11], some of them are directed towards for forecasting solar power [12, 13, 14, 15, 16, 17, 18]. Solar time series data includes high volatility and nonlinearity. Thus, nonlinear time series prediction methods are more preferable to achieve high prediction accuracies in solar power forecasting [8, 10, 14]. Nonlinear methods, such as Artificial Neural Networks (ANNs), Gradient Boosting Machines (GBM), Support Vector Regressor (SVR) can arbitrarily fit complex non-stationary functions with high accuracy.

Multi-scale decomposition techniques are also used together with prediction methods to increase prediction performance. Empirical Mode Decomposition (EMD) is one of the most widely used decomposition technique in different time series problems [9, 19, 20]. EMD is able to decompose any nonlinear and nonstationary data into components [21]. Since these components are highly correlated in themselves, more accurate forecasting results can be achieved. Solar power data usually shows nonlinear and non-stationary properties and has different volatility and fluctuations in different time periods. In order to achieve high accuracy results in solar power forecasting, several methods using EMD are proposed [22, 23, 24].

In this study, Turkey's solar power production is predicted with an ensemble method using EMD which does not only use historical data, but also enriched explanatory feature set which yields better approximations to more accurate predictions. In this ensemble method, the predictions are performed for each EMD components separately and their results are aggregated as final prediction result. Solar energy production is directly proportional with the amount of time that the solar power panels are exposed to the sun. Therefore, during the summer season at noon hours, the highest production figures appear. In order to capture these seasonal variations, the developed ensemble method uses chronological features such as month and quarter. In addition, the angle of the sun and irradiance at ground level play an important role in energy production. Therefore, the developed ensemble method takes various irradiation-related features into account.

The developed method is tested with ANN, SVR, GBM methods in the prediction stage. We performed day-ahead prediction where we predict next day (24-hours) production for every hour. We followed two approaches: Hour-based and Day-based. In Hour-based approach, each hour prediction in a day is performed separately by training the model with the corresponding hours in the historical data. On the other hand, in Day-based approach, each hour prediction in a day is performed by training the model with the entire historical data. We showed that Hour-based approach is more advantageous in solar power predictions, since it captures the high correlation between the production and solar irradiance.

METHODS

The developed method includes three main parts: Multiscale decomposition, feature enrichment and modeling. The architecture of the method is shown in Fig. 1.

Multiscale Decomposition

Solar power data shows nonlinear and non-stationary properties and has different volatility and fluctuations in different time periods. Using multiscale decomposition techniques are efficient and effective method to analyze this nonlinearities in these time series [25]. In literature, there are several multiscale decomposition techniques such as Empirical Mode Decomposition (EMD), Wavelet Packet Decomposition (WPD), Fourier Transform (FT) and etc. [26]. EMD is able to preserve time scale of the given time series throughout the decomposition, thus it is a more preferable technique than the others [21].

The essence of the EMD technique is to empirically identify the intrinsic oscillatory modes in the data and then decompose the data accordingly. The decomposed components, are called Intrinsic Mode Function (IMF) and have two important properties:

1. Each IMF has its own local characteristic time scale.
2. IMFs are relatively stationary series.

These properties allow to resolve nonlinearity and nonstationarity in solar power data. Thanks to this, forecasting methods are able to produce more accurate results.

Let y(t) be a solar power data and EMD technique can be described as follows:

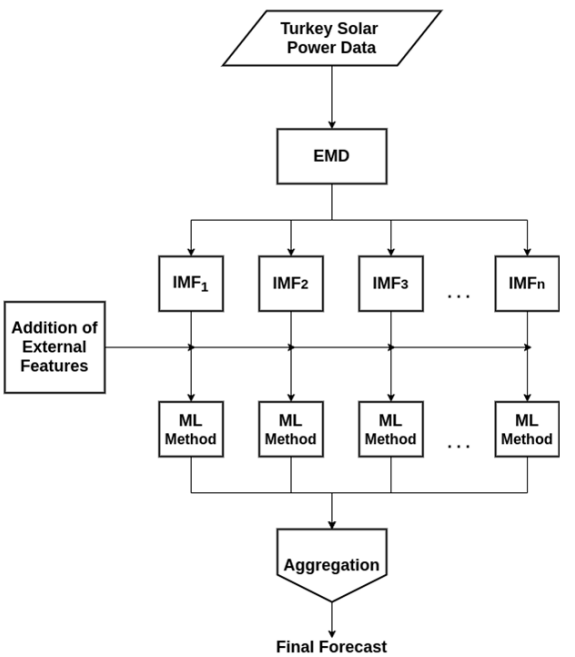


Figure 1. The architecture of proposed ensemble method

$$y(t) = \sum_{k=1}^{m} imf(t)_k + r(t)_m \quad (1)$$

where *imf(t)* represents the intrinsic mode functions, and *r(t)* is the residue. The detailed EMD procedure is as follows [27]:

1. All local extrema maxima and minima points of time series *y(t)* are identified.

2. Upper and lower envelopes are constructed for identified extrema points by using cubic spline interpolation.

3. Mean value m(t) of the envelopes are computed.

4. Take difference of mean and given series and obtain the first components: *h(t)=y(t)-m(t)*

5. Check properties of being IMF for *h(t):*

a. If *h(t)* meets two above conditions, it is considered as an IMF and it is denoted as *imf(t)* in Equation 1. Then, let residual *r(t)=y(t)-h(t)* and replace *y(t)* with the residual *r(t)* return step 1.

b. If *h(t)* does not satisfy the conditions, let *y(t)=r(t)* and return step 1.

6. Steps 1-5 are repeated until the obtained residual *r(t)* is a monotonic function.

Feature Enrichment

The production of solar energy highly depends on various factors such as the presence of the sun, the angle of the sun and irradiance [1, 28]. Therefore, rather than using only historical data, those features can also be included into the feature set to increase forecasting accuracies. These features can be divided into two groups: Chronological and irradiance-related features.

Seasonal variations of solar energy production can be captured with chronological features. For this purpose, month and quarter knowledge is extracted from the time information and added into the feature set.

The monthly distribution of solar energy production is given in Fig. 2. Production values reaches the highest levels in August, but it is very low in December due to insufficient sunny days.

The quarterly distribution of solar energy production is given in Fig 3. It is evident that the production in the fourth quarter of the year, which includes sunless months, is considerably less than that of the third quarter of the year, which includes sunny months.

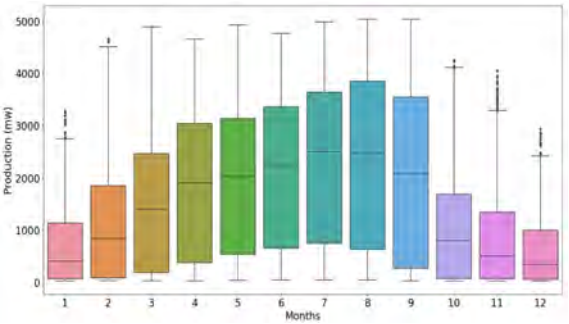


Figure 2. Monthly distribution of the solar power production of Turkey between Sep.2017-Sep.2019

Extraterrestrial, solar-oriented beam irradiance events occurring in the Earth's atmosphere are examined in three different components [1]: Direct Normal Irradiance (DNI), Diffuse Horizontal Irradiance (DHI) and Global Horizontal Irradiance (GHI).

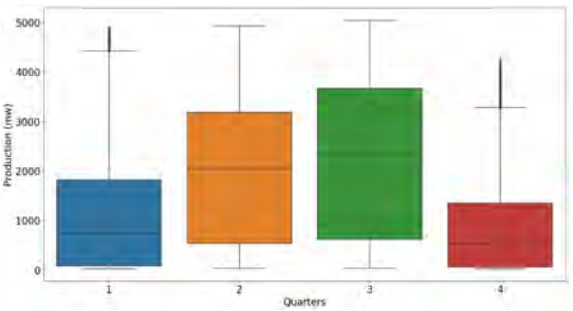


Figure 3. Quarterly distribution of the solar power production of Turkey between Sep.2017-Sep.2019

DNI is the amount of solar radiation that is perpendicular to the Sun, received per unit area at a given location. This measurement is particularly important to solar thermal installations that track the position of the sun. DFI is the amount of solar radiation at the Earth's surface from light scattered by the atmosphere and comes equally from all directions. Unlike DNI, DFI does not take radiation coming directly from the sun into account. GHI is the total amount of irradiance from the sun on Earth. This value is the sum of direct irradiance (DNI) and diffuse irradiance (DHI) which can be shown as follows:

$$GHI = DHI + DNI * \cos q \quad (2)$$

where q is the solar zenith angle. Since GHI indicates the summation of DNI and DFI, solar power plants particularly follow this amount.

These irradiance values, which explain the angle of the sun and quantity of the sun, are important explanatory features for solar energy production. Therefore, these are also added to the attribute set of the developed ensemble method.

Modelling

The solar energy production data *y(t)* is decomposed into n number of components (*imf_1,imf_2,imf_3,...,imf_n*) by EMD, and each component is modelled with a number of the observed the data (imf(t-1),imf(t-2),...,imf(t-a)) and enriched feature set (*f_chronological,f_irradiance*) by machine learning models, as shown in the following equation:

$$y(t+1) = g\left(imf\,(t-1), imf\,(t-2), ..., imf\,(t-a), f_{ch}, f_{irr}\right) \quad (3)$$

where *g* is the machine learning models such as ANN, GBM, SVR methods that we used in this study. Finally, forecasts from each component are summed up.

We propose to model the developed method, which aims to produce day ahead forecast, with an Hour-based approach. In this approach, each hour is trained and modeled in itself. This is because production within each hour is highly correlated in itself. Thus, models trained in data with less variance show higher predictive performance. Fig.4 shows the hourly distribution of solar energy production. While solar power production in the morning and evening hours is low, it increases in noon hours.

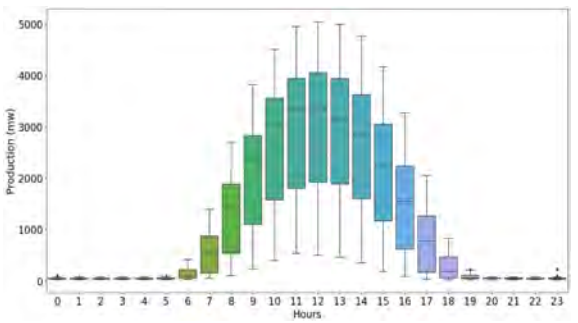


Figure 4. Hourly distribution of the solar power production of Turkey between Sep.2017-Sep.2019

RESULTS AND DISCUSSION

For Turkey solar power forecasting, hourly production data of last two years is used which is shown in Fig.5. This data is publicly available [29]. This hourly dataset consists of 17520 solar production values in Megawatt from Sep. 2017 to Sep. 2019. In this dataset, the solar production values vary between 0 and 5039 Megawatt (MW), while the average hourly production value is 979 MW.

In the experiments, 3 different machine learning algorithms, ANN, GBM, and SVR, are adopted into the developed ensemble method. In the modelling phase, the solar power dataset is split into three sets such as train, validation and test. 25% of the train set is used as the validation set to adjust the hyperparameters of the methods. The last 90 days

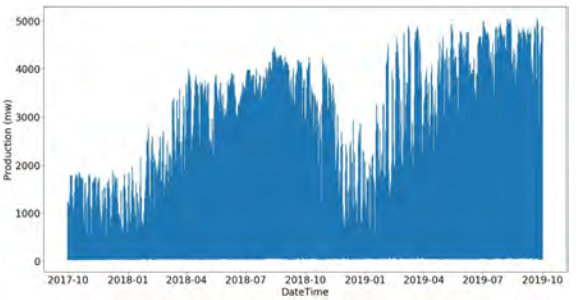


Figure 5. Hourly distribution of the solar power production of Turkey between Sep.2017-Sep.2019

of this time series, which has hourly 720 days of solar power production data, is used as a test set. The last 158 days of the remaining 630 days are the validation set, while 472 days are the training set. The test set used to present the performance of the methods is the last three months of the dataset. We apply grid search to tune the hyperparameters. We provided the search space of each algorithm in Table 1 and we indicated the ones that give the best results in bold. We used those best hyperparameters in order to report our error results in Table 2.

Table 1. The hyper-parameters of the applied machine learning methods

		Parameter search spaces
GBM	•	*learning_rate : [0.01, 0.05, 0.10, 0.20]*
	•	*max_depth : [2, 3, 4, 5]*
	•	*n_estimator : [300, 900, 3900, 4900]*
SVR	•	*kernel = [linear, poly, rbf]*
	•	*C = [0.1, 0.5, 1, 10]*
	•	*epsilon = 0.1*
ANN	•	*activation : [ReLU, Sigmoid, Tanh]*
	•	*optimizer : [Adam, Nadam, Sgd]*
	•	*architecture = (n x 2n x 1)*

In order to show the effectiveness of Hour-based approach, the same experimental setup is implemented for Day-based approach. While comparing performance results between methods, two error metrics are used: Normalized Mean Absolute Error (NMAE) and Normalized Mean Squared Error (NMSE) which are formulated as follows:

$$NMAE = \left(\frac{1}{n}\sum\left|e(t)\right|\right) / \; N$$
$$NMSE = \left(\frac{1}{n}\sum e(t)^2\right) / \; N \quad (4)$$

where *e(t)=y(t)-y^(t)* and *y(t)* is the actual data value, *y^(t)* is the forecasted value at given time t. *N* is the installed solar power capacity. NMAE measures the normalized average of the absolute errors over a given prediction whereas NMSE specifies the normalized average of the squared error.

The experiment results are shown in Table 2. The machine learning methods used in the proposed ensemble architecture are additionally executed individually and their results are also compared in terms of both error metrics. Furthermore, all these experiments are performed for both Day-based and Hour-based approaches, as seen in Table 2.

Table 2. Experimental Results (All NMAE results are multiplied by 100.)

Approach	*Method*	*NMAE*	*NMSE*
Day-based	*GBM*	*2.80*	*14.93*
	SVR	*3.12*	*16.89*
	ANN	*2.72*	*13.00*
	Ensemble – GBM	*2.63*	*12.02*
	Ensemble – SVR	*2.97*	*14.67*
	Ensemble – ANN	*2.71*	*12.88*
Hour-based	*GBM*	*2.55*	*11.41*
	SVR	*3.01*	*15.91*
	ANN	*2.64*	*11.83*
	Ensemble – GBM	*1.96*	*8.07*
	Ensemble – SVR	*2.13*	*10.98*
	Ensemble – ANN	*2.01*	*8.91*

When individual execution of machine learning methods are compared, it is seen that GBM and ANN give better results than SVR. For example, in the Hour-based approach, ANN and GBM give 2.55 and 2.64 NAME error results, respectively, while NAME result of SVR is slightly more than 3. As a result, ANN and GBM are able to yield better generalization in this nonlinear dataset. When the GBM and ANN results are compared in themselves, ANN has better results in individual use in execution, whereas GBM is able to have more accurate results in the proposed ensemble approach.

When the results of the proposed ensemble methods are examined, it is clear that it has improved the results of the individually executed machine learning methods. In both Hour-based and Day-based approaches, ensemble methods are able to produce more accurate results for both error metrics. For example, in the Hour-based approach, GBM, SVR and ANN machine learning methods are improved by the proposed ensemble method by 23%, 28%, and 24%, respectively. This indicates that when machine learning methods are fed with decomposed time series data with descriptive exogenous features, the movement in data can be captured more effectively.

When Day-based and Hour-based approaches are compared, it is clear that the Hour-based approach gives more accurate results than the Day-based approach. As shown in Fig. 4, solar power production values have different distributions within each hour. Hour-based approach, which reduces the variance of these different distributions, outperforms Day-based results between 4% and 25%.

CONCLUSION

Solar power production forecasting plays an important role in Turkey's electricity industry for making decisions in power system scheduling and grid regulation. In this study, we developed an ensemble method which can work with different machine learning algorithms on EMD components of the time series data with explanatory exogenous features. We use this method to forecast solar power production of Turkey. This method firstly decomposes the solar power data into the components by the EMD decomposition technique. These components are then enriched with explanatory exogenous features. Each of these components are modeled separately by machine learning methods and their results are aggregated to obtain final prediction result. We trained our ensemble method with two different approaches, Day-based and Hour-based. We performed the experiments on the last two years of solar energy production data of Turkey. Our experiments show that when (i) the time series data is decomposed before feeding into the machine learning algorithms, (ii) explanatory exogenous features are used in addition to historical data, (iii) each hour in a day is predicted with the model trained with the corresponding hours in the historical data. As a result, the developed ensemble method has a high potential in prediction of solar power production in energy industry.

References

1. Inman RH, Hugo TC, Carlos FM. Solar forecasting methods for renewable energy integration. Volume 39, Issue 6, December 2013, Pages 535–576, doi:10.1016/j.pecs.2013.06.002.
2. Enerji Atlasi, 2019, Gunes Enerji Santralleri, retrieved September 12, 2019, from: https://www.enerjiatlasi.com/gunes/
3. GUYAD, 2019, retrieved October 22, 2019, from: http://www.guyad.org/TR,360/teias-10-yillik-kapasite-planlamasini-yayinladi.html
4. Heinemann D, Lorenz E, Girodo M. Forecasting of solar radiation in: solar energy resource management for electricity generation from local level to global scale. Nova Sciences Pulishers; 2006.
5. Perez R, Beauharnois M, Lorenz E, Pelland S, Schlemmer J. Evaluation of Numerical Weather Prediction Solar Irradiance Forecasts in the US Proc. ASES Annual Conference. Raleigh, NC, USA–1721, May; 2011.
6. Marquez R, Carlos FM. Intra–hour dni forecast based on cloud tracking image analysis.Solar Energy,91, 327–336, May 2013, doi.org/10.1016/j.solener.2012.09.018.
7. Bosh J, Zheng Y, Kleissl J. Deriving cloud velocity from a area of solar radiation measurements. Solar Energy, 87, 196–203, Jan. 2013, doi.org/10.1016/j.solener.2012.10.020.

8. Zhang P. Time Series Forecasting Using a Hybrid ARIMA and NeuralNetwork Model. Neurocomputing 50, 159–175, Neurocomputing, 50, 159-175, 2002, doi:10.1016/S0925–2312(01)00702–0.
9. Buyuksahin UC, Ertekin S. Improving forecasting accuracy of time series data using a new ARIMA–ANN hybrid method and empirical mode decomposition, Neurocomputing 361, 151–163, 2019, doi.org/10.1016/j.neucom.2019.05.099.
10. Khashei M, Bijari M. A Novel Hybridization of Artificial Neural Networksand ARIMA Models for Time Series Forecasting. Appl. Soft Comput., 11, 2664-2675, 2011, doi:10.1016/j.asoc.2010.10.015.
11. Buyuksahin UC, Ertekin S. A feature–based hybrid ARIMA–ANN model for univariate time series forecasting. Journal of the Faculty of Engineering and Architecture of Gazi University 35:1 (2020) 467–478 doi.org/10.17341/gazimmfd.508394.
12. Huang J, Korolkiewicz M, Agrawal M, Boland J. Forecasting solar radiation on an hourly time scale using a coupled autoregressive and dynamical system (cards) model. Solar Energy; 87, 136–149, Jan. 2013, doi.org/10.1016/j.solener.2012.10.012.
13. Glasbey CA, Allcroft DJ. A spatiotemporal auto–regressive moving average model for solar radiation. Appl Stat, 57, 343–355, 2007.
14. Mellit, A. Artificial Intelligence Technique for Modelling and Forecasting of Solar Radiation Data: A Review. Int. J. Artif. Intell. Soft Comput., 1, 52-76, 2008.
15. Martin L, Zarzalejo LF, Polo J, Navarro A, Marchante, R, and Cony M. Prediction of Global Solar Irradiance Based on Time Series Analysis: Application to Solar Thermal Power Plants Energy Production. Solar Energy, 84, 1772–1781, Oct. 2010, doi.org/10.1016/j.solener.2010.07.002.
16. Mohamed A, Chowdhury C. Solar Power Forecasting Using Artificial Neural Networks. In 2015 North American Power Symposium (NAPS), 1-5, 2015.
17. Hamid E, Himdi K. Artificial Neural Network for Forecasting One Day Ahead of Global Solar Irradiance. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, May 29, 2018.
18. Daniel O, Kubby J. Feature Selection and ANN Solar Power Prediction. Research Article. Journal of Renewable Energy, 2017. https://doi.org/10.1155/2017/2437387.
19. Xu X, Qi Y, Hua Z. Forecasting demand of commodities after natural Disasters. Expert Systems with Applications, Volume 37, Issue 6, June 2010, Pages 4313–4317, doi.org/10.1016/j.eswa.2009.11.069.
20. Yu W, Mu–Chen C. Forecasting the short–term metro passenger flow with empirical mode decomposition and neural networks, Transportation Research Part C, Volume 21, Issue 1, April 2012, Pages 148–162, https://doi.org/10.1016/j.trc.2011.06.009.
21. Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non–stationary time series analysis. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 454 (1998) 903–995. doi:10.1098/rspa.1998.0193.
22. Barnhart BL, Eichinger WE. Empirical mode decomposition applied to solar irradiance, global temperature,sun spot number and co2 concentration data. J Atmos Solar Terr Phys 2011; 73:1771.
23. Majumder I, Behera MK, Nayak N. Solar Power Forecasting Using a Hybrid EMD–ELM Method. International Conference on Circuit, Power and Computing Technologies (ICCPCT), 20–21 April 2017, doi:10.1109/ICCPCT.2017.8074179.
24. Monjoly S, Andre M, Calif R, Soubdhan T. Hourly forecasting of global solar radiation based on multiscale decomposition methods: A hybrid approach. Energy, Volume 119, 15 January 2017, Pages 288–298, doi.org/10.1016/j.energy.2016.11.061.
25. Calif R, Schimtt FG, Huang Y, Soubdhan T. Intermittency study of high frequancy global solar raiation sequences under a tropical climate. Solar Energy, 98, 349–365, 2013.
26. Huang NE, Wu ML, Qu W, Long SR, Shen SP. Applications of Hilbert Huang transform to non–stationary nancial time series analysis. Applied Stochastic Models in Business and Industry, 19(3), (2003), 245–268.
27. Angela Z, Faltermeier R, Keck I, Tomé A, Puntonet C, Lang E. Empirical Mode Decomposition – an Introduction. Proceedings of the International Joint Conference on Neural Networks, 1-8, 2010.
28. Kutlu, C, Li J, Su Y, Wang Y, Pei G, Riffat S. Annual Performance Simulation of a Solar Cogeneration Plant with Sensible Heat Storage to Provide Electricity Demand for a Small Community: A Transient Model. Hittite Journal of Science & Engineering, 6(1), (2010), 75–81.
29. Epias, 2018, Gerçek Zamanlı Üretim, retrieved October 26, 2018, from: https://seffaflik.epias.com.tr/transparency/uretim/gerceklesen–uretim/gercek–zamanli–uretim.xhtml