HAL
open science

Accurate determination of parameters relationship for photovoltaic power output by augmented dickey fuller test and engle granger method

Harry Ramenah, Philippe Casin, Moustapha Ba, Michel Benne, Camel Tanougast

To cite this version:

Harry Ramenah, Philippe Casin, Moustapha Ba, Michel Benne, Camel Tanougast. Accurate determination of parameters relationship for photovoltaic power output by augmented dickey fuller test and engle granger method. AIMS Energy, 2018, AIMS Energy, 6 (1), pp.19-48. 10.3934/energy.2018.1.19 . hal-01697126

HAL Id: hal-01697126

https://hal.univ-reunion.fr/hal-01697126v1

Submitted on 30 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AIMS Energy

AIMS Energy, 6(1): 19–48.
DOI: 10.3934/energy.2018.1.19
Received: 27 October 2017
Accepted: 19 December 2017
Published: 02 January 2018

http://www.aimspress.com/journal/energy

Research article

Accurate determination of parameters relationship for photovoltaic power output by augmented dickey fuller test and engle granger method

Harry Ramenah[1,*], Philippe Casin[1], Moustapha Ba[1], Michel Benne[2] and Camel Tanougast[1]

[1] LCOMS Laboratory, University of Lorraine, 7 Rue Marconi, 57070 Metz, France
[2] LE2P Laboratory, University of Reunion, 40 Avenue René Cassin, 97400 Saint-Denis, France

* **Correspondence:** Email: harry.ramenah@univ-lorraine.fr; Tel: +33372749325; Fax: +33387315257.

Abstract: Power output from photovoltaic (PV) systems in outdoor conditions is substantially influenced by climatic parameters such as solar irradiance and temperature. PV manufacturers always provide technical specifications in laboratory conditions but reliable relationship for the power output must be determined for accurate prediction under real operating conditions. For the present study, solar irradiance G, temperature T and electrical power output P data under real conditions are methodically and regularly inscribed in dataloggers. Hence, in this paper, we investigate rigorous and robust statistical methods for small sample such as Augmented Dickey-Fuller and Engle Granger for stationary series to determine the estimate regression between variables P, G & T. A first regression of power output P time series variable on solar irradiance G time series has shown spurious results and thus spurious regression. The first differences of such time series are stationary and a regression is proposed whereas temperature variable is identified as not significant and where autocorrelation of residuals is suspected. Finally, the novelty of this paper is the Engle & Granger method that is used to provide a relationship between variables P and G in a difference level. A final relationship without suspicious heteroscedasticity has been determined. Our model is formulated on the basis of PV real conditions statistical approach and is more realistic than steady approach models.

Keywords: photovoltaic; unit root; Augmented Dickey Fuller; correlograms; Engle Granger test; difference series; spurious regression; heteroscedasticity

1. Introduction

Forecasting accurately photovoltaic (PV) power output is not only based on climatic conditions but large variation of the power output is also observed due to several factors such as solar irradiance, temperature, wind speed and humidity. Actually PV nominal specifications such as power output or energy yield of recent modules are evaluated by manufacturers under standard conditions (STC) by a flashing technology and these nominal specifications are hand out to customers. Under real outdoor conditions theses specifications are not always valid. Meanwhile, an economical solar simulator based on micro-channel solar cell thermal (MCSCT) [1–3] has been designed and tested in indoor condition. Yet, the PV measured data from indoor and outdoor conditions are decorrelated. Since late 70 s the simplest explicit equation [4] for the steady-state operating temperature of a solar module links with the ambient temperature and the incident solar radiation flux had been studied. It is difficult to effectively assess the impact of PV output variability [5,6] on the power grid stability without a clear understanding of the factors that influence this variability.

Many factors have become redundant in the presence of other factors, and their value in the forecasting framework may vary under different prediction horizons and error measurements. Nevertheless, many studies conducted to date have shown that temperature [7–9] is the most important parameter that influences the PV output, although the latter is largely dependent on solar irradiance. Comparative studies of different models have been used to predict the PV module temperature in a mathematically [10] explicit and several empirical models [6,11,12] for estimation of PV module temperature have been proposed based on solid state physics of PV or thermal radiation and convection contributions [13–15] to the PV operating temperature as well as for PV technologies such as thin film PV system under Mediterranean climate conditions [16]. More implicit in form or experimental results have validated temperature effect on electrical model [17] in outdoor conditions. In the past few years, many statistical techniques [18–21] have also been proposed to forecast energy output of PV system or others have shown either how to improve confidence intervals of PV data from estimator's variance [22] or modeling solar forecasting through Artificial Neural Network [23–26] as an alternative to conventional approaches. Also a novel contribution of fault detection algorithm [27] has been proposed using a statistical analysis of real-time long term measured data and theoretical thresholds.

Some studies have used classical regression methods that took advantage of correlation nature of meteorological variables which are used prediction model as inputs. However, regression between non-stationary series may lead to conclude on the existence of a relationship between two variables even though there is no meaningful linear relationship between them. The novelty of this paper is to show that if PV explanatory variables are non stationary then a study of first difference must be determined. An incidental advantage of the first-difference transformation is that it may make a

nonstationary time series stationary to reach the final PV equation taking into account the temporal relations between the PV explanatory variables in a moderate zone.

We compared measurements and prediction of energy output [28] in real outdoor conditions as a function of location and environmental conditions mainly solar irradiance φ (W/m^2). The goal of the present work carried out is to develop an operational forecasting framework zone. An accurate PV output forecasting method would be of great help to grid operators and would minimize costs while enabling more integration of variable renewable energy (RE) in the grid. The resulting statistical model should then be applied in future works on collected datasets at several particular temperate climatic sites for validation. A mathematical model [29] using satellite data has also been proposed to determine PV performance in continental scale. The two models should be compared in a future work.

Statistically we studied a linear relation analysis of time series data collected over a period of time and investigated mainly the dependent variable of power output P on explanatory variables such as solar irradiance G and temperature T. For that we first determined the stationary character or not of each time series of solar irradiance and temperature as this will set the statistical method to be used. Testing the stationarity of a series would be able to understand autocorrelation function and its statistical significance. Dickey Fuller (DF) have developed a test called Augmented Dickey & Fuller (ADF) [30,31] which is a unit root test for stationarity.

The ADF test is used to determine the method of regression estimation between PV variables such as P, G and T. Then we show that these variables are not stationary at level and an ordinary least square (OLS) regression is only possible at difference level for each variable including a constant in the regression. First results of the study show that temperature at first difference level and the constant are not very significant as the OLS coefficients are well estimated. Then we analyze residual hypotheses.

Analysis of residuals is a powerful diagnostic tool and it helps to assess whether some of underlying assumptions of regression have been violated. Ideally all residuals should be small and unstructured meaning that the regression analysis has been successful in explaining the essential part variation of the dependent variable. We first apply the Goldfeld-Quandt (GQ) test when heteroscedastic variance is related to variables in the regression model then the Durbin Watson (DW) test to detect serial correlation in the residuals. In this present study, the GQ test is well verified but DW test is still doubtful and not very conclusive. Outliers are suspected in the model. Therefore, a more recent procedure such as the Engle & Granger (EG) method is suggested to determine the most appropriate model. Our model is formulated on the basis of PV real conditions statistical approach and is more realistic than steady approach models.

For this study, the small samples of one year daily means data is retrieved among the 7 years measurements from the GREEN lab of Physics department of University of Lorraine in Metz. The PV design of the GREEN lab is an on grid connected system. Six polycrystalline modules of SCHÜCO technologies are connected in a series wiring pattern and mounted on the south-south east vertical wall of the platform building. Each module has a peak power of 205 W_P, at tilt angle of 60 °, low ventilation and connected to a SCHÜCO inverter for a power level up to 1 kW_P.

This paper is organized as follows. Section 2 is an introduction to the methodology explaining the ADF stationary test and Engle Granger tests for cointegration. In Section 3, we describe the input

solar irradiance distribution of this peculiar zone to predetermine if a frequency distribution such as Weibull distribution is expected. In Section 4, we use informal tools such correlograms and Q-statistic to test whether each variable such as solar irradiance G, power output P and temperature T is stationary or non-stationary. The ADF test is applied to each variable and results are presented and analyzed in table form. If a time series has a unit root then first differences of such time series are assume to become stationary. In that section, we apply the ADF test to first difference series transforming non-stationary time series to yield stationary series, we propose a probable estimates difference equation regression model. In section 5, we propose the OLS regression equation and identify outliers before analyzing residuals from the propose regression using the GQ test then the DW-d stat test with its corresponding tabulated bound. We discuss and highlight the proposed regression equation with residuals and introduction to EG method. In section 6, we apply the EG correction mechanism reconciling the short-run behavior with its long-run behavior and a final relationship without suspicious heteroscedasticity is proposed. Finally, conclusion and future work are indicated in Section 7.

2. Methodology

Let us consider a dependent variable Y and a number of explanatory variables such as X's (X_1, X_2, ..., X_k). X_k, being the kth explanatory variable. These (k+1) variables will denote the ith observation on explanatory variable X_k. In order to determine the linear relationship between variables, for example, $Y = a_0 + a_1X_1 + \ldots + a_jX_j + \ldots + a_kX_k$, the well-known least squares linear regression technique is used. This technique provides an estimate of each of the coefficients for a_i, giving a measure of the quality of the linear fit of the correlation coefficient R and provides confidence intervals for the predictions, etc. The interpretation of R in a multiple regression model is of dubious value and this technique can only be used when observations of one explanatory variable is independent one to each other.

However, and in this study, this hypothesis of independency is not necessarily verified, due to time series: there exists a relationship between the y(t) series and y(t–1), y(t–2) series and so on, there may be an autocorrelation phenomenon for X_k variables as well.

In this context, it is then necessary to check whether the series are stationary or not. Indeed, performing a regression between non-stationary series leads to what is known as spurious results or regression [29] for R^2 determination as well as for the regression coefficients. This can lead to conclude the existence of a relationship between two variables when in fact there is no linear relationship between them. Regressions involving time series data include the possibility of obtaining spurious or dubious results [29] in the sense that superficially results look good but on further probing they look suspect.

If variables are non stationary which is the case of this study, then the first difference (or upper orders if needed) must be studied. If these first differences are stationary then the multiple regression technique is used to determine linear relationships between theses first differences. An incidental advantage of the first-difference transformation is that it may make a nonstationary time series stationary.

A linear combination of variables at a non stationary level may become stationary which is given as the co-integration equation also known as the Engel Granger (EG) method [32,33]. Highlighting of such a relationship should allow improving the estimation of the linear relationship between the variables in difference that should also lead to reach the final equation, first by taking into account the temporal relations between the variables, then by introducing explanatory variables as variables measured in the previous period.

2.1. Dickey-Fuller stationary test

The principle of DF [9] in statistics stationary test is based on null hypothesis or test for the existence of a unit root in an autoregressive model. Unit root test is explained in appendix A. DF have shown that under the null hypothesis that is if $\pi = 0$ then the estimated t value of the coefficient of X_{t-1} follows the τ (tau) statistic. The tau statistic or test is known as the DF test. DF has computed critical tau values and can be obtained in a tabulated form.

The DF test is usually estimated on three models with their corresponding equation under three different null hypotheses. But we should consider only one random walk equation for the model as others are mainly used by economists in a financial field.

Model: X_t is a random walk, where there is no constant or drift component and no trend:

$$X_t = \rho X_{t-1} + u_t \quad \text{or} \quad \Delta X_t = \pi X_{t-1} + u_t \tag{1}$$

In each case, the null hypothesis is given for $\pi = 0$, that is there is a unit root and the time series is non-stationary. The alternative hypothesis is that π is less than zero and the time series is stationary.

In this paper, the following indication for different hypotheses will be used:

Hypothesis H_0 $\pi = 0$ means $\rho = 1$ then such series is a non-stationary series.

Hypothesis H_1 $\pi < 0$ means $\rho < 1$ then such series is a stationary series.

It is extremely important to note that the critical values of the tau test to test the hypothesis that $\pi = 0$, are different for each of the preceding three specifications of the DF test. The estimation procedure of tau will be explained later from experimental data table. However, if the computed absolute value of the tau statistic (τ) less than the DF critical tau values from tables, hypothesis that $\pi = 0$ is rejected in which case the time series is stationary. On the other hand, if the computed τ is less than the critical tau value the null hypothesis is not rejected in which case the time series is non-stationary.

In conducting the DF test it was assumed that the white noise error term u_t was uncorrelated. But in case the u_t are correlated then DF have developed a test, known as the Augmented Dickey-Fuller (ADF) test which is explained in the next section.

2.2. Augmented Dickey-Fuller stationary test

To tackle autocorrelation problems DF have developed a test called Augmented Dickey-Fuller (ADF) which is still a unit root test for stationarity. This test is conducted by augmenting the

above equation by adding the lagged values of the dependent variable ΔX_t. This equation is given as follows:

$$\text{Model (1): } X_t = \rho X_{t-1} + \sum_{j=1}^{p} \Psi_j \Delta X_{t-j} + u_t \text{ or } \Delta X_t = \pi X_{t-1} + \sum_{j=1}^{p} \Psi_j \Delta X_{t-j} + u_t \quad (2)$$

Where $\Delta X_{t-1} = (X_{t-1} - X_{t-2}), \Delta X_{t-2} = (X_{t-2} - X_{t-3})$, etc.

The number of lagged difference terms to include is often determined empirically, the goal is to include enough terms so that the error is serially uncorrelated. In ADF still test whether $\pi = 0$ and the ADF test follows the same asymptotic distribution as the DF statistic so the same critical values can be used.

2.3. Spurious regression

Let us consider a random walk model given by Eq 3a:

$$Z_t = Z_{t-1} + v_t \text{ and } X_t = X_{t-1} + u_t \quad (3a)$$

And the same number of observations generated from v_t et u_t and also assumed that the initial values of both Z and X were zero. Il is also assumed that u_t and v_t are serially uncorrelated as well as mutually uncorrelated. Suppose Z_t is regressed on X_t as both are uncorrelated processes the R^2 from the regression of Z on X should be close to zero; that is, there should not be any relationship between the two variables. However, it may happen that the R^2 value is significant but there is no linear relationship between the variables. This is the phenomenon of spurious regression and spurious model is not desirable [33].

To avoid spurious regression problem that may arise from regressing a non-stationary time series on one or more non-stationary time series, we have to transform non-stationary time series to make them stationary. If a linear combination of non-stationary variables is stationary, the variables are said to be cointegrated [33]. Their fluctuations are concomitant. But the regression model is not appropriate for highlighting linear relationships between non-stationary variables, this is the problem of spurious regression, so other procedures such as Engle & Granger must be used as described in the next section.

2.4. Engle-Granger tests for cointegration

A linear combination of non stationary variables can be stationary, the variables are then cointegrated and their fluctuations are related. But the OLS regression model is not appropriate for highlighting linear relationships between non stationary variables, this is the problem of spurious regression as discussed earlier. Another procedure such as EG [34,35] is then applied. The EG test for cointegration is a three-step procedure.

The first step: it is necessary to ensure that the first differences of the corresponding Z_t and X_t series are stationary series.

The second step: let v_t be the residual of the regression of Z_t with respect to X_t given as follows:

$$Z_t = aX_t + b + v_t \tag{3b}$$

then if v_t is stationary then Z_t and X_t series are cointegreted and the relationship is usually called long run equilibrium. Practically it's not possible to do a stationary test on the unknown z_t series but the test can be applied to the following series:

$\hat{v}_t = Z_t - \hat{a}\, X_t - \hat{b}$ where hat a and hat b are determined by regressing Z_t with respect to X_t. The critical values of DF test and ADF test are then modified.

The third step: Error Correction Model (ECM)

The model is as follows:

$$\Delta Z_t = \gamma \hat{v}_{t-1} + \beta \Delta X_{t-1} + \delta + \varepsilon_t \text{ and } \Delta Z_t = \gamma \left(Z_{t-1} - \hat{a}\, X_{t-1} - \hat{b} \right) + \beta \Delta X_{t-1} + \delta + \varepsilon_t \tag{3c}$$

The three variables of this equation ΔZ_t, γ_{t-1} and ΔX_{t-1} are stationary so that all estimations of this equation are given by the least squares method. If γ is a negative value, v_{t-1} acts as spring force and thus when v_{t-1} is a positive value the Z_t value has a higher value than it should be given the long run relationship, $Z_{t-1} > aX_{t-1} + b$. As γ has negative value the v_{t-1} effect on ΔZ_t gives a negative value and thus $Z_t < Z_{t-1}$ that tends to balance the long run relationship. For those interested on that topic may refer to a more specialized books in econometrics.

3. Solar irradiance frequency distribution test

Solar energy is harnessed with solar PV converting sunlight directly into electricity. Thus the energy output of PV system depends on the input solar irradiance distribution. In this section, the solar data of a peculiar site is systematically recorded and analyzed to determine whether the distribution of solar irradiance frequency distribution follows a predetermined distribution such as Weibull distribution which is then used to predict performance and energy output as for wind turbine [34]. Although 8 years of solar irradiance data have been recorded from the GREEN Platform of University of Lorraine and due to data similarities between years, only one year data is used for the test. Otherwise the 8 years data that is being sampled every 10 minutes would need huge processing time.

These long term irradiance data are used to calculate the probability density function of the irradiance for different hours of a typical day in a month. The frequency distribution of solar irradiation of the study series is displayed as a histogram in Figure 1. The histogram divides the series range (the distance between the maximum and minimum values) into a number of equal length intervals or bins and displays a count of the number of observations that fall into each bin. Figure 1 indicates the number of observation against 15 bins of classes and each class is an interval of values of the irradiance variable G (W/m^2). For example, the fourth bin between 75 to 100 indicate the corresponding number of 17 observations. The visual examination does not seem to identify any appropriate functional form of a Weibull distribution curve that may be superimposed on the histogram pattern. We therefore proceed by a statistical test as indicated in Table 1.

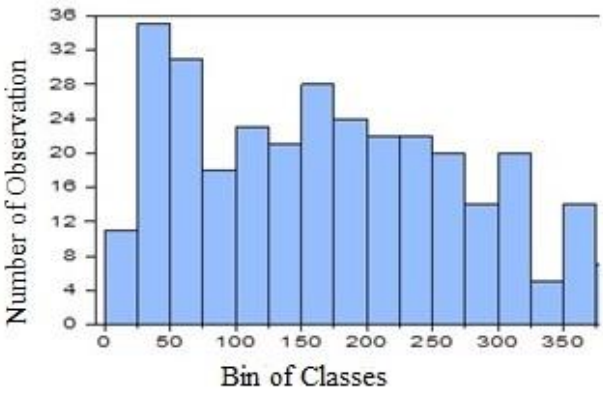

Figure 1. Rejecting the normal distribution of solar irradiance.

The statistical hypothesis testing framework is as follows, the null hypothesis H_0 is Weibull distribution and H_1 is non Weibull distribution. The statistical test is based on the likelihood ratio that gives the number of times data is more likely under one model than the other.

This likelihood ratio is then used to determine the probability or p-value or compared to a critical value to decide whether or not to reject the null hypothesis. Table 2 is the second part of the output of Table 4. The p-value is less than the required significance level then we say the null hypothesis is rejected at the given level of significance. But at the same time the z-statistic value is so great compared to test critical values.

Therefore, the null hypothesis of a Weibull distribution cannot be retained but seems to be very unlikely.

Table 1. Empirical distribution of solar irradiance test.

Method	Value	Adj. Value	Probability
Cramer-von Mises	0.159053	0.160722	(0.01, 0.025)
Watson	0.157972	0.159630	(0.01, 0.025)
Anderson-Darling	1.250903	1.264034	< 0.01

Hypothesis Weibull; Sample: 363; Observation: 363.

Table 2. The likelihood table.

Parameter	Value	Std. Error	z-statistic	Probability
M	0.000000	-	NA	NA
S	234.2153	8.624198	27.15477	0.0000
A	1.500020	0.062977	23.81856	0.0000
Log likelihood	–2267.794	Mean dependent var		211.5373
No. of coefficients	2	S.D dependent var		141.4750

4. From pattern recognition to stationary assumption

Figure 2 represents 363 observations of the solar irradiance G evolution to demonstrate the application of ADF test on variable.

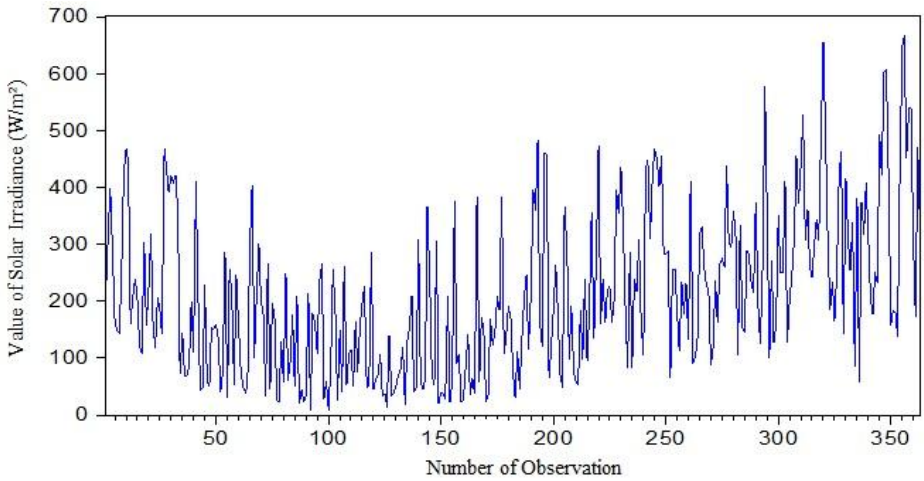

Figure 2. Irradiance evolution for year observation.

Before we go to the ADF model, we draw the graph of the study series for an eye examination of any pattern that might be important in our future assumption. In the next sections, we should verify whether the series are stationary or not.

4.1. Correlograms & Q-Statistics

In that part the two alternative hypothesis are considered, H_0: series X is stationary; H_1: series X is not stationary.

We shall be using correlograms or Ljung box statistics to test whether the series is stationary or not. The following tables show pictures of correlograms of a time series. Autocorrelation and partial autocorrelation functions characterize the pattern of temporal dependence in the series and typically make sense only for time series data. This is applied to solar irradiance data and is displayed in Table 3. The first column is the autocorrelation column represented as spikes between vertical lines.

The autocorrelation of a series Z at lag k is estimated by the following Eq 4:

$$\tau_k = \frac{\sum_{t=k+1}^{T}(Z_t-\bar{Z})(Z_{t-k}-\bar{Z})}{\sum_{t=1}^{T}(Z_t-\bar{Z})} \tag{4}$$

where $\bar{Z}$ is the sample mean of Z. This is the correlation coefficient for values of the series k periods apart. If τ1 is non zero, it means that the series is first order serially correlated. If τk dies off more or less geometrically with increasing lag k it is a sign that the series obeys a low-order autoregressive (AR) process. If τk drops to zero after a small number of lags it is a sign that the series obeys a low order moving-average (MA) process.

Normally, when the spikes are outside the two lines we suspect that the data is not stationary and this is for example one sign among others. From the AC column in Table 3, the estimated values at lag k that are usually noted as hat symbol $\hat{\rho}_k$ is gradually going down which means that probably data is non-stationary and is lagged from 1 to 36.

Table 3. Correlogram solar irradiance series.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.498	0.498	90.872	0.000
		2	0.365	0.155	139.74	0.000
		3	0.355	0.169	186.19	0.000
		4	0.278	0.031	214.64	0.000
		5	0.218	0.012	232.26	0.000
		6	0.214	0.051	249.31	0.000
		7	0.244	0.102	271.53	0.000
		8	0.336	0.202	313.56	0.000
		9	0.318	0.073	351.43	0.000
		10	0.301	0.051	385.46	0.000
		11	0.295	0.032	418.28	0.000
		12	0.259	0.008	443.66	0.000
		13	0.225	0.012	462.83	0.000
		14	0.205	0.013	478.82	0.000
		15	0.194	0.012	493.19	0.000
		16	0.236	0.066	514.54	0.000
		17	0.275	0.082	543.41	0.000
		18	0.260	0.020	569.35	0.000
		19	0.236	-0.016	590.87	0.000
		20	0.244	0.021	613.84	0.000
		21	0.231	0.020	634.55	0.000
		22	0.166	-0.047	645.20	0.000
		23	0.150	-0.011	653.95	0.000
		24	0.233	0.110	675.16	0.000
		25	0.252	0.060	700.12	0.000
		26	0.234	0.012	721.59	0.000
		27	0.271	0.056	750.56	0.000
		28	0.289	0.044	783.52	0.000
		29	0.271	0.030	812.75	0.000
		30	0.225	-0.004	832.83	0.000
		31	0.188	-0.013	846.99	0.000
		32	0.177	-0.019	859.57	0.000
		33	0.181	0.004	872.73	0.000
		34	0.190	0.019	887.24	0.000
		35	0.207	0.013	904.53	0.000
		36	0.258	0.065	931.47	0.000

The partial autocorrelation at lag k is the regression coefficient on Z_{t-k} when Z_t is regressed on a constant and Z_{t-1}, … , Z_{t-k}. This is a partial correlation since it measures the correlation of Z values that are k periods apart after removing the correlation from the intervening lags. When the pattern of autocorrelation is one that can be captured by an autoregression of order less than k then the partial autocorrelation at lag k will be close to zero that is indicated by the PAC (Partial AutoCorrelation) column in Table 3.

Finally, the last value of the Q-stat (Q-Statistic) column in Table 3 are significant at all lags indicating significant serial correlation in the residuals. The last value of the Q-stat column which 931.47 is a high value and the corresponding probability or p-value is zero which less that 5% meaning that the null hypothesis H_0 is rejected.

We note that equation for Q-stat is given as follows:

$$Q = T(T+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{T-K} \tag{5}$$

where T is the total number of observations et m is the lag length.

The stationary test of the solar irradiance series seems to be a non-stationary data so OLS between these series could not be applied. The slow linear decay of the AC coefficients in Table 3 can be observed which indicates the need to differentiate. An advantage of the first-difference transformation is that it may make a non stationary time series stationary.

Correlogram of the first difference for solar irradiance is given in Table 4, the autocorrelations at various lags hover around zero which is a picture of the correlogram of a stationary time series. We repeated the procedure for power output and temperature data. Tables 5 & 7 respectively, show each correlogram of power and temperature series with the corresponding correlogram of the first difference series given in Tables 6 & 8. These correlograms show similarities to the solar irradiance series. The null hypothesis is thus rejected and all series are non-stationary series but not for the first difference series.

The ADF is discussed in the next section to determine the functional form of the regression model with the first difference as variables.

Table 4. Correlogram first difference solar.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.370	-0.370	50.000	0.000
		2	-0.119	-0.297	55.185	0.000
		3	0.066	-0.128	56.780	0.000
		4	-0.022	-0.100	56.952	0.000
		5	-0.055	-0.130	58.088	0.000
		6	-0.032	-0.165	58.478	0.000
		7	-0.065	-0.253	60.027	0.000
		8	0.109	-0.117	64.418	0.000
		9	0.003	-0.087	64.422	0.000
		10	-0.007	-0.060	64.439	0.000
		11	0.030	-0.033	64.772	0.000
		12	-0.002	-0.034	64.773	0.000
		13	-0.014	-0.033	64.847	0.000
		14	-0.013	-0.035	64.909	0.000
		15	-0.056	-0.091	66.097	0.000
		16	0.004	-0.106	66.104	0.000
		17	0.055	-0.039	67.269	0.000
		18	0.008	-0.003	67.292	0.000
		19	-0.027	-0.037	67.572	0.000
		20	0.020	-0.038	67.722	0.000
		21	0.054	0.032	68.855	0.000
		22	-0.050	-0.003	69.830	0.000
		23	-0.100	-0.125	73.751	0.000
		24	0.062	-0.076	75.254	0.000
		25	0.039	-0.023	75.863	0.000
		26	-0.057	-0.068	77.145	0.000
		27	0.025	-0.052	77.382	0.000
		28	0.030	-0.044	77.735	0.000
		29	0.034	-0.004	78.195	0.000
		30	-0.014	-0.002	78.278	0.000
		31	-0.024	0.008	78.502	0.000
		32	-0.016	-0.012	78.600	0.000
		33	-0.006	-0.027	78.616	0.000
		34	-0.004	-0.015	78.623	0.000
		35	-0.039	-0.074	79.225	0.000
		36	0.023	-0.068	79.433	0.000

Table 5. Correlogram of power series.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.436	0.436	69.564	0.000
		2	0.324	0.166	108.14	0.000
		3	0.333	0.181	148.91	0.000
		4	0.262	0.052	174.20	0.000
		5	0.218	0.034	191.81	0.000
		6	0.198	0.031	206.42	0.000
		7	0.196	0.051	220.78	0.000
		8	0.302	0.195	254.89	0.000
		9	0.279	0.077	284.01	0.000
		10	0.257	0.052	308.76	0.000
		11	0.258	0.036	333.88	0.000
		12	0.255	0.046	358.51	0.000
		13	0.192	-0.027	372.49	0.000
		14	0.187	0.020	385.78	0.000
		15	0.176	0.012	397.53	0.000
		16	0.217	0.070	415.48	0.000
		17	0.238	0.066	437.19	0.000
		18	0.250	0.064	461.17	0.000
		19	0.224	-0.001	480.53	0.000
		20	0.224	0.007	499.96	0.000
		21	0.245	0.062	523.21	0.000
		22	0.157	-0.065	532.82	0.000
		23	0.151	0.002	541.73	0.000
		24	0.251	0.129	566.39	0.000
		25	0.239	0.055	588.73	0.000
		26	0.212	-0.001	606.42	0.000
		27	0.238	0.034	628.74	0.000
		28	0.272	0.067	657.92	0.000
		29	0.254	0.014	683.42	0.000
		30	0.204	-0.009	700.05	0.000
		31	0.165	-0.019	710.98	0.000
		32	0.146	-0.053	719.49	0.000
		33	0.150	-0.015	728.53	0.000
		34	0.142	0.000	736.59	0.000
		35	0.145	-0.011	745.10	0.000
		36	0.204	0.046	761.95	0.000

Table 6. Correlogram of difference power series.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.403	-0.403	59.385	0.000
		2	-0.103	-0.317	63.277	0.000
		3	0.069	-0.151	65.034	0.000
		4	-0.026	-0.120	65.292	0.000
		5	-0.021	-0.107	65.458	0.000
		6	-0.017	-0.120	65.564	0.000
		7	-0.096	-0.248	68.989	0.000
		8	0.115	-0.121	73.900	0.000
		9	0.001	-0.088	73.901	0.000
		10	-0.018	-0.054	74.021	0.000
		11	0.004	-0.071	74.028	0.000
		12	0.053	0.004	75.090	0.000
		13	-0.051	-0.040	76.064	0.000
		14	0.002	-0.035	76.066	0.000
		15	-0.048	-0.093	76.945	0.000
		16	0.017	-0.089	77.060	0.000
		17	0.010	-0.083	77.102	0.000
		18	0.032	-0.019	77.503	0.000
		19	-0.020	-0.022	77.659	0.000
		20	-0.019	-0.078	77.795	0.000
		21	0.097	0.050	81.468	0.000
		22	-0.073	-0.015	83.513	0.000
		23	-0.096	-0.144	87.106	0.000
		24	0.098	-0.073	90.821	0.000
		25	0.014	-0.013	90.902	0.000
		26	-0.046	-0.047	91.749	0.000
		27	-0.003	-0.075	91.753	0.000
		28	0.042	-0.026	92.439	0.000
		29	0.031	0.001	92.823	0.000
		30	-0.012	0.006	92.880	0.000
		31	-0.016	0.043	92.984	0.000
		32	-0.022	0.008	93.175	0.000
		33	0.010	-0.007	93.216	0.000
		34	-0.007	0.011	93.238	0.000
		35	-0.053	-0.052	94.361	0.000
		36	0.008	-0.096	94.385	0.000

Table 7. Correlogram of temperature series.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.828	0.828	251.12	0.000
		2	0.742	0.178	453.14	0.000
		3	0.686	0.110	626.37	0.000
		4	0.636	0.049	775.75	0.000
		5	0.598	0.052	908.22	0.000
		6	0.596	0.139	1040.2	0.000
		7	0.606	0.130	1176.8	0.000
		8	0.626	0.141	1323.0	0.000
		9	0.612	-0.000	1463.2	0.000
		10	0.600	0.028	1598.1	0.000
		11	0.594	0.056	1730.7	0.000
		12	0.577	0.022	1856.4	0.000
		13	0.549	-0.019	1970.4	0.000
		14	0.545	0.051	2083.3	0.000
		15	0.552	0.065	2199.4	0.000
		16	0.563	0.064	2320.3	0.000
		17	0.574	0.063	2446.4	0.000
		18	0.568	-0.005	2570.5	0.000
		19	0.560	0.005	2691.3	0.000
		20	0.537	-0.032	2802.6	0.000
		21	0.512	-0.013	2904.1	0.000
		22	0.494	-0.008	2998.7	0.000
		23	0.488	0.015	3091.6	0.000
		24	0.513	0.109	3194.3	0.000
		25	0.525	0.027	3302.3	0.000
		26	0.532	0.019	3413.5	0.000
		27	0.540	0.029	3528.2	0.000
		28	0.555	0.084	3650.2	0.000
		29	0.537	-0.022	3764.5	0.000
		30	0.516	-0.016	3870.3	0.000
		31	0.488	-0.051	3965.5	0.000
		32	0.480	0.005	4057.6	0.000
		33	0.483	0.041	4151.2	0.000
		34	0.470	-0.045	4240.3	0.000
		35	0.484	0.045	4335.1	0.000
		36	0.503	0.037	4437.7	0.000

Table 8. Correlogram of difference temperature.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.259	-0.259	24.481	0.000
		2	-0.076	-0.154	26.603	0.000
		3	-0.027	-0.098	26.863	0.000
		4	-0.055	-0.113	27.962	0.000
		5	-0.108	-0.192	32.266	0.000
		6	-0.036	-0.179	32.746	0.000
		7	-0.032	-0.188	33.124	0.000
		8	0.104	-0.045	37.158	0.000
		9	0.014	-0.052	37.235	0.000
		10	-0.002	-0.070	37.236	0.000
		11	0.032	-0.036	37.619	0.000
		12	0.028	0.005	37.915	0.000
		13	-0.070	-0.056	39.778	0.000
		14	-0.048	-0.087	40.646	0.000
		15	-0.021	-0.091	40.815	0.000
		16	-0.002	-0.090	40.817	0.001
		17	0.058	-0.015	42.123	0.001
		18	0.005	-0.034	42.132	0.001
		19	0.061	0.020	43.553	0.001
		20	0.011	0.008	43.599	0.002
		21	-0.012	0.014	43.651	0.003
		22	-0.043	-0.009	44.365	0.003
		23	-0.102	-0.116	48.425	0.001
		24	0.019	-0.055	48.572	0.002
		25	0.031	-0.020	48.940	0.003
		26	-0.001	-0.030	48.940	0.004
		27	-0.007	-0.080	48.957	0.006
		28	0.084	-0.008	51.720	0.004
		29	0.025	0.011	51.960	0.006
		30	0.000	0.029	51.960	0.008
		31	-0.051	-0.009	52.990	0.008
		32	-0.041	-0.043	53.663	0.010
		33	0.049	0.051	54.621	0.010
		34	-0.068	-0.024	56.487	0.009
		35	-0.030	-0.050	56.856	0.011
		36	0.055	-0.034	58.085	0.011

4.2. Augmented dickey fuller solar irradiance test

To tackle the autocorrelation problem, the Augmented Dickey-Fuller (ADF) is applied to test the null hypothesis of whether a unit root unit is present in a time series test. The test is applied to the three series solar irradiance series (G), power output series (P) and temperature series (T) with the following assumption, H_0: series has a unit root; H_1: series has no unit root.

The ADF test results are divided into two distinct sections. The first portion displays the test of the unit root output provides information about the form of the test that is, the type of test, the exogenous variables and lag length used and contains the test output associated critical values such as the probability p-value. The second part of the output shows the intermediate test equation computed to calculate the ADF statistic. In this section, the unit root test is applied to the solar irradiance and the results are explained. In the following sections, the same procedure is then applied to the power and temperature series.

The first portion of the ADF test is given in Table 8a, where the none exogenous estimate the test that does not include a constant and linear trend in the test regression and the variable is E that is lagged once as displayed by the correlogram of Table 3.

Table 8a. Testing solar irradiance.

		t-Statistic	Probability*
Augmented Dickey-Fuller test statistic		–0.778035	0.3786
Test critical values:	1% level	–2.571511	
	5% level	–1.941721	
	10% level	–1.616099	

Null Hypothesis: G has a unit root; Exogenous: None; Lag Length: 8 (Automatic: based on SIC, maxlag = 16).

ADF statistic absolute value in Table 8a is –0.778035 and the associated one-sided p-value is –0.3786 for a number of observations specified in Table 8b. In addition, the critical values at the 1%, 5% and 10% levels are also reported. We notice here that the t-statistic value is greater than the critical values which is higher than 5% so that we do not reject the null hypothesis at conventional test sizes.

Second part of the output is displayed in Table 8b and shows the intermediate test equation that has been used to calculate the ADF statistic upon 363 observations and the dependent variable is D (G) which is regressed first lag.

In Table 8b, the column labeled coefficient depicts the estimated coefficients or estimates and should not be viewed as being deterministic. It's a kind of indication upon the variable precision. The absolute t-statistic value associated to the G (–1), that is 0.778035 is defined as the ratio coefficient to the standard error and given as follows:

$$\frac{Coefficient\ value}{Std.Error\ value} = \frac{\widehat{m}}{\hat{\sigma}_m} = \frac{-0.021473}{0.027599} = -0.778035$$

The t-stat value of the coefficient value should be compared to critical values tabulated by DF. This value is less than those indicated in Table 2, given as 2.58, 1.95, 1.61 for the corresponding

threshold 1%. This comparison indicates that the solar irradiance series is not stationary. In this test we ignored the probability value from Table 8b because the coefficient value of G (−1) is not distributed that is it does not follow the student distribution. The t-statistic value and corresponding probability value of the first difference term lagged one are highly significant so we did not make any mistake from earlier analysis in considering this variable.

In the following sections, only results of the first part of the ADF test corresponding to power output and temperature data are presented as the second parts have similar characteristics to the first difference lagged at different levels to the second part of solar irradiance ADF test.

Table 8b. Second output of ADF solar equation.

Variable	Coefficient	Std. Error	t-Statistic	Probability
E (–1)	–0.021473	0.027599	–0.778035	0.4371
D (E (–1))	–0.632395	0.058947	–10.72823	0.0000
D (E (–2))	–0.525921	0.064797	–8.116418	0.0000
D (E (–3))	–0.354483	0.067566	–5.246454	0.0000
D (E (–4))	–0.328680	0.067350	–4.880199	0.0000
D (E (–5))	–0.366181	0.066697	–5.490251	0.0000
D (E (–6))	–0.374641	0.066079	–5.669598	0.0000
D (E (–7))	–0.337516	0.062371	–5.411434	0.0000
D (E (–8))	–0.124641	0.054153	–2.301633	0.0220
R-squared	0.335576	Mean dependent var		–0.290650
Adjusted R-squared	0.320169	S.D. dependent var		142.5432
S.E. of regression	117.5296	Akaike info criterion		12.39635
Sum squared resid	4765554.	Schwarz criterion		12.49472
Log likelihood	–2185.154	Hannan-Quinn criter		12.43549
Durbin-Watson stat	2.028671			

Augmented Dickey-Fuller Test Equation; Dependent Variable: D (E); Method: Least Squares; Sample: (adjusted 363); Included observations: 354 after adjustments.

4.3. Augmented dickey fuller power & temperature test

Similar ADF test is applied to the variable power (P) and temperature (T) and each first part is given in appendix C. Examination of these tables reveals that the observed p-value in each case given respectively as 0.3459 and 0.6907 is not significant as it is a very high value compare to 5%. Also t-stat value of the variable power (P) equal to −0.853089 is again higher than all the critical values. The t value of the temperature coefficient is 0.026823 but this value is greater than even the 10 percent critical τ value of –1.616101 suggesting that even after taking care of possible autocorrelation in the error term, both power and temperature series is a non-stationary series. However, the both dependent variable D (P) and D (T) first difference lagged one shows significant characteristics values of a stationary series.

The next section is concerned with the ADF test applied to difference series.

4.4. ADF test to difference series of irradiance, power and temperature

To avoid the spurious regression problem that may arise from regressing a non stationary time series on one or more non stationary time series, we have to transform non stationary time series to make them stationary. If a time series has a unit root then the first differences of such time series may be stationary. Therefore, the solution here is to take the first differences of the time series.

Table 9a. First part difference solar irradiance G.

		t-Statistic	Probability*
Augmented Dickey-Fuller test statistic		–12.22837	0.0000
Test critical values:	1% level	–2.571511	
	5% level	–1.941721	
	10% level	–1.616099	

Null Hypothesis: D (G) has a unit root; Exogenous: None; Lag length: 7 (Automatic-based on SIC, mawlag = 16).

Table 9b. Second part difference solar irradiance G.

Variable	Coefficient	Std. Error	t-Statistic	Probability
D (G (–1))	–4.145076	0.338972	–12.22837	0.0000
D (G (–1), 2)	2.492967	0.311017	8.015532	0.0000
D (G (–2), 2)	1.949729	0.275106	7.087194	0.0000
D (G (–3), 2)	1.579856	0.234643	6.733011	0.0000
D (G (–4), 2)	1.237359	0.192906	6.414317	0.0000
D (G (–5), 2)	0.859116	0.149072	5.763101	0.0000
D (G (–6), 2)	0.474450	0.101222	4.687212	0.0000
D (G (–7), 2)	0.129271	0.053794	2.403044	0.0168
R-squared	0.758759	Mean dependent var	–0.900650	
Adjusted R-squared	0.753878	S.D. dependent var	236.7688	
S.E. of regression	117.4625	Akaike info criterion	12.39246	
Sum squared resid	4773916	Schwarz criterion	12.47990	
Log likelihood	–2185.465	Hannan-Quinn criter	12.42725	
Durbin-Watson stat	2.029534			

Augmented Dickey-Fuller Test Equation; Dependent Variable: D (G, 2); Method: Least Squares ; Sample: (adjusted 363); Included observations: 354 after adjustments.

D operator is used to specify differences of series. To specify first differencing, we simply include the series name in parentheses after D with the corresponding lagged. The number of lagged difference terms to include is often determined empirically, the idea being to include enough terms so that the white noise error term in is serially uncorrelated. The difference solar irradiance table of the ADF test is illustrated in following tables. The similar tables of each dependent variable P and T are given in appendix C.

The lag length is being checked with the corresponding correlogram of the series to determine and to re-estimate with one less round of differencing.

4.5. Conclusion

The time series of variable P, G, T are not stationary series but differencing may yield stationary series. We can therefore use the regression test between the 3 variables in difference. This is discussed in the next section.

5. Difference regression equation

We computed the regression as each variable is stationary in difference. Table 10, is the regression difference where the explanatory variables ΔG and ΔT are used including a constant C. The dependent variable is ΔP (Delta P) for 363 samples and 362 included observations after adjustments. Let's just turn to interpreting the results. From Table 14, firstly we deduce the expected regression equation referring to the corresponding explanatory coefficient.

This is given as Eq 6a below:

$$\Delta P = 0.969\,\Delta G + 0.500\,\Delta T\ - 0.105 \tag{6a}$$

Table 10. The regression difference data.

Variable	Coefficient	Std. Error	t-Statistic	Probability
ΔG	0.969868	0.027980	34.66231	0.0000
ΔT	0.500517	0.603774	0.828980	0.4077
C	–0.010523	1.778432	–0.005917	0.9953
R-squared	0.945447	Mean dependent Var	0.284254	
Adjusted R-Squared	0.945143	S.D dependent var	144.4618	
S.E of regression	33.83531	Akaike Info criterion	9.889140	
Sum squared resid	410993.4	Schwarz criterion	9.921391	
Log Likelihood	–1786.934	Hannan-Quinn criter	9.901961	
F-statistic	3110.855	Durbin-Watson stat	2.799248	
Prob (F-Statistic)	0.000000			

Dependent Variable: ΔP; Method: Least Squares; Sample: (adjusted 363); Included observations 362 after adjustments.

It is a standard practice to use the coefficient p-values to decide whether to include variables in the final model. Yet, the corresponding p-value of ΔT and constant term C are not statistically significant because their corresponding p-value (0.4077, 0.9953) are greater than the usual significance level of 1%, 5%, 10% and can be removed from Eq 6a. Whereas the t-statistic value of ΔG is highly significant as well as the corresponding probability that is less than 5%, expecting good regression between ΔP and ΔG.

The overall regression fit as measured by the R^2 value is more than 94% indicating a very tight fit. Obviously we'll focus on R-squared (R^2) suggesting that solar irradiance alone can explain over nearly 95% of the variation of power output and it will have only 1% impact on the variable power P. Although this regression seems to be very significant however, both the constant and temperature difference are not significant as is revealed by the corresponding probability p-value 0.99 and 0.40. Indeed, the p-value given just below the F-statistic in Table 10 denoted as Prob (F-statistic**)** is the

marginal significance level of the F-test. If the p-value is less than the significance level that is 5% the hypothesis that all slope coefficients are equal to zero are rejected.

For example, in Table 10, the p-value is essentially zero for ΔG. Note that the F-test is a joint test so that even if all the t-statistics are insignificant the F-statistic can be highly significant. However, we can propose the apparent estimate equation using least squares given as follows by Eq 6b.

$$\Delta P = 0.969\ \Delta G \tag{6b}$$

The proposed regression equation is still not satisfactory because other data in the table seem to indicate differences between the predicted value (based on the regression equation) and the actual observed value. This is discussed in the next section.

5.1. Outliers & analyzing residuals from regression

One problem with least squares occurs when there are one or more large deviations for example, cases whose values differ substantially from the other observations. These points are called outliers. They may represent important information about the relationship between variables. Robust regression might be a good strategy since it is a compromise between excluding these points entirely from the analysis and including all the data points and treating all them equally in ordinary least squared (OLS) regression. In linear regression, an outlier is an observation with large residual.

Figure 3, represents the residual graph where the blue curve is the residuals from the regression provided by a residual table. The green and red curves are respectively the real or actual and fitted curves of power output against the number of observations. We see that the regression seems to be going rather well from the point of view of predicting power output evolution. However, the residual graph shows 4 outliers diagnosing failures of the above assumptions of the regression model. The 4 outliers do appear on one side at 220 and 221 observations with amplitude around 100 and on the other side at 357 and 358 with amplitude around 170.

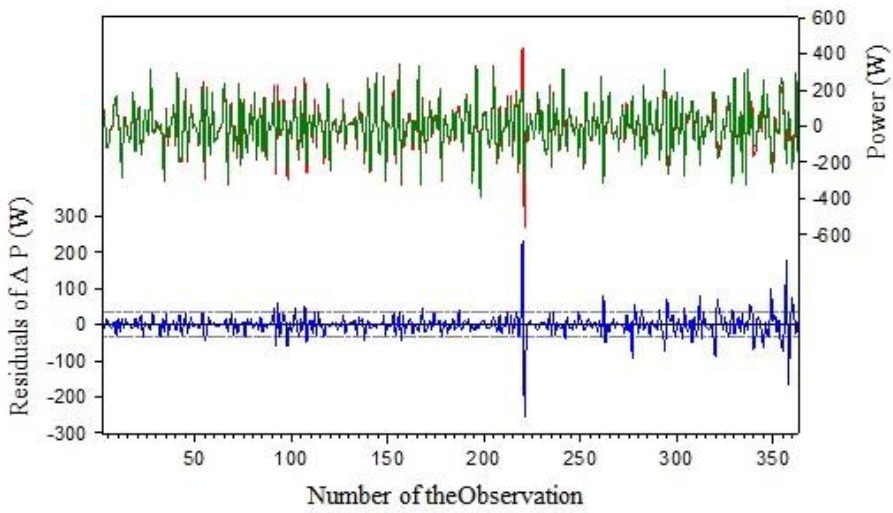

Figure 3. The real & fitted curves and residuals graph.

Given the existence of outliers, we re-estimate the regression without these 4 points as given in Table 11 with the corresponding residual curve of ΔP as illustrated in Figure 4.

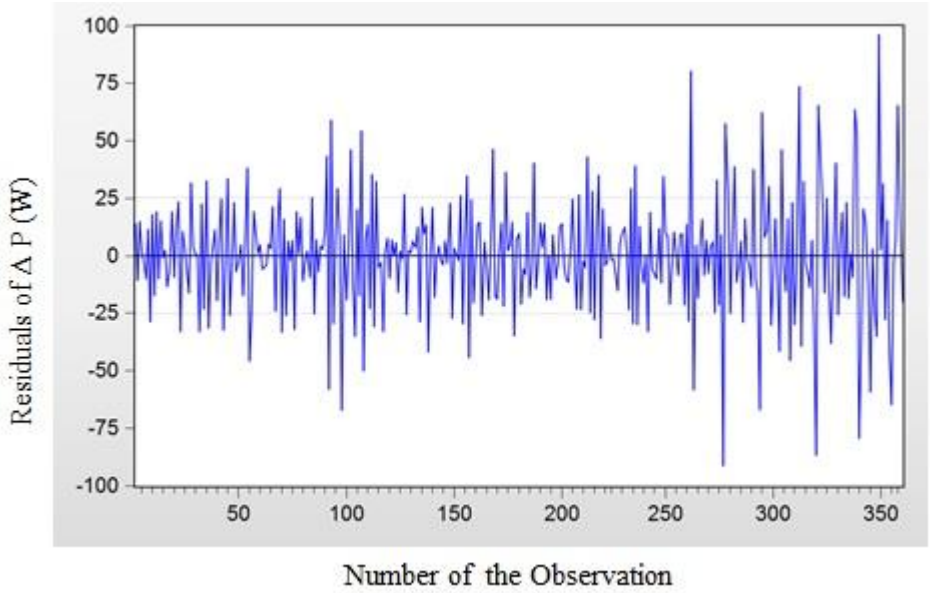

Figure 4. The residual graph of ΔP.

From the residual graph of ΔP in Figure 4, we can see that there is no longer any spurious residual but we can suspect a little heteroscedasticity, that is to say a variance of the residual values varying with the level of the variables. This is discussed in the next section.

Table 11. Regression difference data without outliers.

Variable	Coefficient	Std. Error	t-Statistic	Probability
C	0.024771	1.338098	0.018512	0.9852
ΔG	0.996329	0.021519	46.29970	0.0000
ΔT	–0.389609	0.461854	–0.843577	0.3995
R-squared	0.967559	Mean dependent Var	0.949972	
Adjusted R-Squared	0.967376	S.D dependent var	140.1690	
S.E of regression	25.31743	Akaike Info criterion	9.309208	
Sum squared resid	227545.1	Schwarz criterion	9.341726	
Log Likelihood	–1663.348	Hannan-Quinn criter	9.322140	
F-statistic	5293.958	Durbin-Watson stat	2.732836	
Prob (F-Statistic)	0.000000			

Dependent Variable: ΔP; Method: Least Squares; Sample: (adjusted 359); Included observations: 358.

5.2. Heteroscedasticity & autocorrelation

When using some statistical techniques such as OLS, a number of assumptions are typically made. In a linear regression model in which the errors have expectation zero and are uncorrelated as

well as having equal variances then the best linear unbiased estimator (BLUE) of the coefficients is given by the OLS estimator. If the variance of the error term ε_i is not the same across all observations i=1, ..., n, then the disturbances are said to be heteroscedastic.

Heteroscedasticity tends to produce p-values that are smaller than they should be. This effect occurs because heteroscedasticity increases the variance of the coefficient estimates but the OLS procedure does not detect this increase. Consequently, OLS calculates the t-values and F-values using an underestimated amount of variance.

This problem can lead to conclude that a model term is statistically significant when it is actually not significant as the coefficients are biased. In our case, as the t-test for each coefficient examines them individually the t-test of ΔG in Table 11 is higher than that in Table 10 with the F-value very significant in Table 11. The overall F-test is significant and the R-squared value has been improved therefore correlation between the model and dependent variable is statistically significant.

We thus identified the single significant explanatory variable which is ΔG with heteroscedastic disturbance and therefore the kind of heteroscedasticity must be identified. In the next section we propose more in-depth studies in order to propose a regression equation closer to real experimental conditions.

The corresponding graph of first difference of power output ΔP against first difference of solar irradiance ΔG is illustrated in Figure 5. However, some scattered values do not fit the regression line reasonably and although the constant term has been removed, the very slight intercept on the delta P axis are indications to pursue further studies.

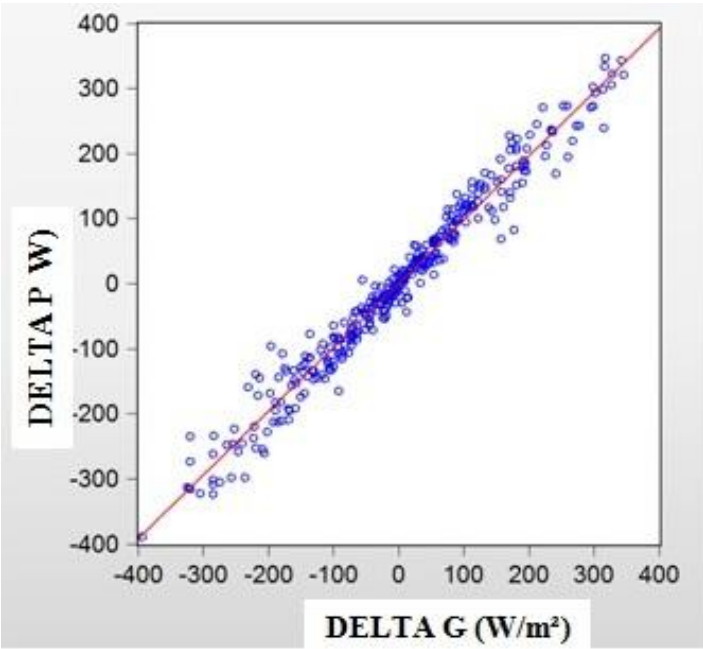

Figure 5. Relation between ΔP and ΔG.

Table 12. Final regression difference data.

Variable	Coefficient	Std. Error	t-Statistic	Probability
ΔG	0.980052	0.009507	103.0827	0.0000
S.E of regression	25.27172	Akaike Info criterion	9.300038	
Sum squared resid	228001.5	Schwarz criterion	9.310878	
Log Likelihood	–1663.707	Hannan-Quinn criter	9.304349	
F-statistic	10595.79	Durbin-Watson stat	2.722380	
Prob (F-Statistic)	0.000000	Mean dependent var	0.949972	

Dependent Variable: ΔP; Method: Least Squares; Sample: (adjusted 359); Included observations: 358.

We then test that the residual is an increasing function of the explanatory variable and the Goldfeld Quandt (GQ) test is proposed. This is discussed in the next section.

5.3. The goldfeld quandt test

The blue curve of Figure 6 is the residual of ΔP with no outliers and is oscillating in a range of values that act as good estimates. This range is also known as the confidence interval and is represented by the dotted blue lines. GQ have argued that the error term may not satisfy the ordinary least squares assumptions and may itself be heteroscedastic.

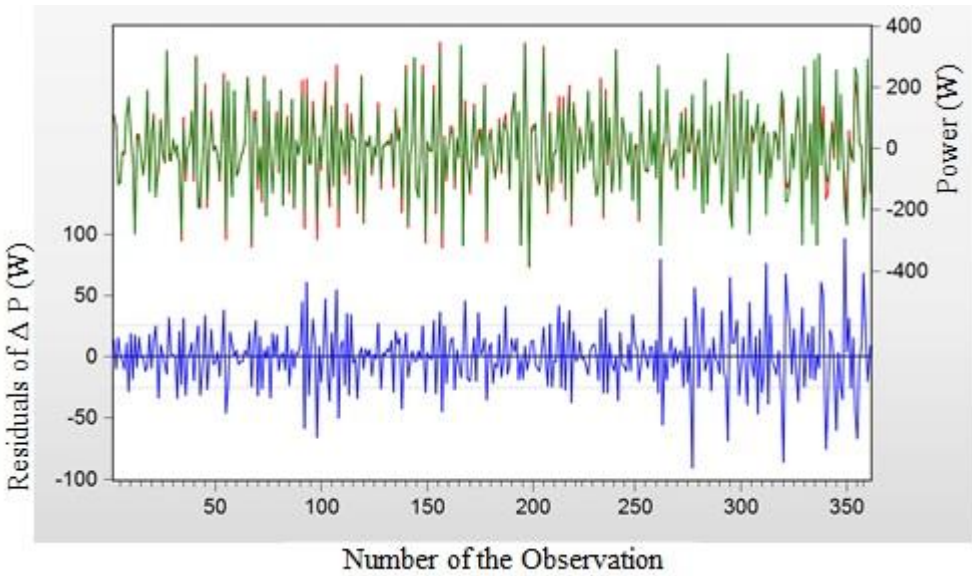

Figure 6. The residual graph of ΔP with no outliers.

The GQ test is accomplished by undertaking separate least squares analyses on two subsets of the original dataset: these subsets are specified so that the observations for which the pre-identified explanatory variable takes the lowest values are in one subset, with higher values in the other. The test statistic or F-test used is the ratio of the mean square residual (MSR) errors for the regressions

on the two subsets given by Eq 7 where MSR_{max} is the highest value in a subset and MSR_{min} is the lowest value.

$$F = \frac{MSR_{max}}{MSR_{min}} \tag{7}$$

In our study, the OLS regressions of the GQ test is based on the first 100 and the last 100 observations and their associated residual sums of squares are given respectively in Tables 13a and 13b.

Table 13a. Regression based on the first 100 observations.

Variable	Coefficient	Std. Error	t-Statistic	Probability
ΔG	0.980428	0.017402	56.33851	0.0000
S.E of regression	31.99997	Akaike Info criterion	9.779297	
Sum squared resid	101375.8	Schwarz criterion	9.805349	
Log Likelihood	–487.9648	Hannan-Quinn criter	9.789840	
Durbin-Watson stat	1.901483	S.D dependent var	75.58948	
Mean dependent var	–166.9105			

Dependent Variable: ΔP; Method: Least Squares; Sample: 260359; Included observations: 100.

Table 13b. Regression based on the last 100 observations.

Variable	Coefficient	Std. Error	t-Statistic	Probability
ΔG	0.978710	0.015881	61.62793	0.0000
S.E of regression	29.54568	Akaike Info criterion	9.619702	
Sum squared resid	86421.76	Schwarz criterion	9.645753	
Log Likelihood	–479.9851	Hannan-Quinn criter	9.630245	
Durbin-Watson stat	1.754791	S.D dependent var	72.19141	
Mean dependent var	169.8802			

Dependent Variable: ΔP; Method: Least Squares; Sample: 260359; Included observations: 100.

From Tables 13a and 13b, the corresponding sum squared residuals values are MSR_{max} = 101375 and MSR_{min} = 86421. The F ratio as indicated above is determined and is equal to 1.17 so we can reject the hypothesis of heteroscedasticity.

However, this value from the GQ method is a non-zero value for the error term and a further study on the DW d-stat test is proposed in the next section.

5.4. Durbin Watson D-Stat test

The corresponding residual graph is displayed in Figure 6, the colored curves have the same defintion as defined in Figure 3 where the blue curve indicates that no outliers values are blatant for the residuals but does not explain the DW statistic value. The conventional DW tables are not applicable when a constant term does not exist in the regression and instead an appropriate set of Durbin-Watson tables is used as reference table.

The DW is a test that the residuals from a linear regression or multiple regression are independent and makes it possible to detect a first order autocorrelation errors that is an equation given as follows:

$$u_t = \rho\, u_{t-1} \quad + \quad \eta_t \tag{8}$$

Because most regression problems involving time series data exhibit positive autocorrelation the hypotheses usually considered in the DW test are H_0: ρ = 0 and H_1: ρ > 0.

The test statistic is given as follows:

$$d = \frac{\sum_{t=2}^{T} (u_t - u_{t-1})^2}{\sum_{t=2}^{T} \left(u_t^2\right)} \tag{9}$$

For $u_t = y_t - \hat{y}_t$ where y and hat y are respectively, the observed and predicted values of the response variable for individual t. d becomes smaller as the serial correlations increase. Upper and lower critical values d_2 and d_1 have been tabulated for different values of the number of explanatory variables and T. If a test for negative autocorrelation is desired, then the 4-d statistic is used. The decision rules for H_0, ρ = 0 versus H_1 ρ < 0 are the same as those used in testing for positive autocorrelation.

The DW table at 5% level d_1 = 1.664 and d_2 = 1.779. The tabulated bound is summarized as follows:

$0 < d < d_1$: positive autocorrelation of residuals

$d_1 < d < 4\text{-}d_2$: no autocorrelation

$4\text{-}d_2 < d < 4\text{-}d_1$: undefined

$4\text{-}d_1 < d < 4$: negative autocorrelation of residuals

The DW statistic from Table 16 is 2.72 which is between 2.336 and 4 therefore a negative autocorrelation is expected.

However, a more rigorous method such as Cochrane Orcutt method [29] can be applied if required but this is beyond the scope of this study.

5.5. Discussion

Regression of a non stationary series on another non stationary series may cause spurious regression which is not desirable. In our case, we have two variables P and G and they have unit root at level meaning non stationary. They are stationary after first difference. Nevertheless the regression model is not appropriate for highlighting linear relationships between non stationary variables and suppose P is regressed on G as follows:

$$P_t = \gamma G_t + C + e_t \tag{10}$$

where e_t has a unit root and is stationary as explained in the previous section. It is then proposed to estimate the regression of equation (10) and obtain the residuals.

However, there is one precaution to exercise. Since the estimated e_t are based on the estimated cointegrating parameter γ, ADF critical significance values are not quite appropriate and other values are used and determined from Engle and Granger method. This is discussed in the next section.

6. Engle and granger method

The outline of Engle & Granger (EG) [30] method is explained in section 2.4, still we can simply resume the principle. EG note that a linear combination of two or more series may be stationary, in which case we say the series are cointegrated. Such a linear combination defines a cointegrating equation between variables with cointegrating vector of weights characterizing the long-run relationship between the variables. Therefore, from Eq 10 is deduced the following relationship:

$$e_t = P_t - \gamma G_t - C \tag{11a}$$

This is computed and the results are given in distinct tables.

Table 14a shows result of regression between variables at level, that is long term relationship. The dependent variable is P and the explanatory variables are solar irradiance G and a constant term C.

Table 14a. Long term data for regression.

Variable	Coefficient	Std. Error	t-Statistic	Probability
G	0.93982	0.010494	89.55454	0.0000
C	16.67559	2.669565	6.246557	0.0000
R-squared	0.956926	Adjusted R-squared	0.956807	
S.E of regression	28.24849	Akaike Info criterion	9.525451	
Sum squared resid	228069.7	Adjusted R-squared	0.956807	
		Akaike Info criterion	9.525451	
Log Likelihood	–1726.869	Schwarz criterion	9.546908	
F-statistic	8020.015	Hannan-Quinn criter	9.533980	
Prob (F-Statistic)	0.000000	Durbin-Watson stat	1.494379	
		Mean dependent var	215.4845	
		S.D dependent var	135.9218	

Dependent Variable: P; Method: Least Squares; Sample: 363; Included observations: 363.

When the dependent variable P_t is regressed on G_t the following regression is obtained

$$P_t = 0{,}9398G_t + 16{,}6755 + e_t \tag{11b}$$

So we focus on stationarity or not of the residual of Table 14c. The residual is noted as ResidCoint.

Since the computed t value (–14.677) is very significant and much more negative than –2.5713 [30,33,36], our conclusion is that the residuals from the regression of P on G are stationary.

Hence, Eq 17b is a cointegrating equation and is called the long run function and interprets its parameters as long run parameters. Thus, 0.9398 represents the long-run or equilibrium. Obviously, in the short run there may be disequilibrium. Therefore, the error term in Eq 11a must be processed as the equilibrium error and this error term is used to link the short-run behavior of PCE to its long-run value.

Table 14b. EG residcoint data.

Variable	Coefficient	Std. Error	t-Statistic	Probability
ResidCoint (–1)	–0.747317	0.050917	–14.67710	0.0000
R-squared	0.373717	Mean dependent var	0.0181680	0.0000
Adjusted R-squared	0.373717	S.D. dependent var	34.53230	
S.E. of regression	27.32818	Akaike info criterion	9.4564721	
Sum squared resid	269605.4	Schwarz criterion	9.4672335	
Log likelihood	–1710.622	Hannan-Quinn criter	9.4607460	
Durbin-Watson stat	2.045377			

Dependent Variable: D (ResidCoint); Method: Least Squares; Sample: (adjusted 363); Included observations: 362 after adjustments.

Table 14c. EG critical values.

		t-Statistic	Probability*
Augmented Dickey-Fuller test statistic		–14.67710	0.0000
Test critical values:	1% level	–2.571336	
	5% level	–1.941701	
	10% level	–1.616113	

Null Hypothesis: ResidCoint has a unit root; Exogenous: None; Lag Length: 0 (Automatic-based on SIC, maxlag = 16).

The error correction mechanism (ECM) which is the third step of the EG test as described in section 2.4 is applied and the similar Eq to (3b) is given as follows:

$$\Delta P_t = \alpha_0 + \alpha\Delta G_t + \beta e_{t-1} + \varepsilon_t \tag{12}$$

Where, Δ as usual denotes the first difference operator, ε_t is a random error term, and

$$e_{t-1} = (P_{t-1} - \gamma G_{t-1} - C) \tag{13}$$

that is the one lagged value of the error from the cointegrating regression of Eq 11b.

ECM equation, that is Eq 12, states that ΔP depends on ΔG and also on the equilibrium error term e_{t-1}. If the latter is non zero, then the model is out of equilibrium. Suppose ΔG is zero and e_{t-1} is positive this means ΔP_{t-1} is too high to be in equilibrium. Since β is expected to be negative the term βe_{t-1} is negative and therefore ΔP_t will be negative to restore the equilibrium. That is, if P_t is above its equilibrium value it will start falling in the next period to correct the equilibrium error. As well as

if e_{t-1} is negative that is P is below its equilibrium value βe_{t-1} will be positive which will cause ΔP_t to be positive leading P_t to rise in period t. Thus, the absolute value of β decides how quickly the equilibrium is restored.

Eq 13 is computed the corresponding results are given in Table 15. The coefficient β tells us at what rate it corrects the previous period disequilibrium of the system. When β is significant and contains a negative sign it validates that there exists a long run equilibrium relationship among the variables.

Table 15. ECM regression data.

Variable	Coefficient	Std. Error	t-Statistic	Probability
ResidCoint (–1)	–0.725527	0.051027	–14.21850	0.00000
ΔG	0.969314	0.010166	95.35067	0.00000
C	0.013603	1.423750	0.009554	0.99244
R-squared	0.965033	Mean dependent var	0.284254	
Adjusted R-squared	0.964836	S.D. dependent var	144.4618	
S.E. of regression	27.08863	Akaike info criterion	9.444358	
Sum squared resid	263432	Schwarz criterion	9.476610	
Log likelihood	–1706.429	Hannan-Quinn criter	9.457179	
Durbin-Watson stat	2.045261	F-statistic	4953.946	
Prob (F-statistic)	0.000000			

Dependent Variable: ΔP; Method: Least Squares; Sample: (adjusted 363); Included observations: 362 after adjustments.

The coefficient value β of ResidCoint (–1) or (e_{t-1}) = –0.7255 that validate the long run relationship between P and G.

Therefore, from Table 19 we can deduce that the model is very significant as the R squared value is very high thus the Fisher statistic value is high. The two explanatory variables that is, first difference stationary solar irradiance ΔG, ResidCont lagged one (e_{t-1}) are very significant whereas the constant term C = α_0 is not significant as indicated by its p-value.

Still, we plot the residual graph and this is displayed in Figure 7.

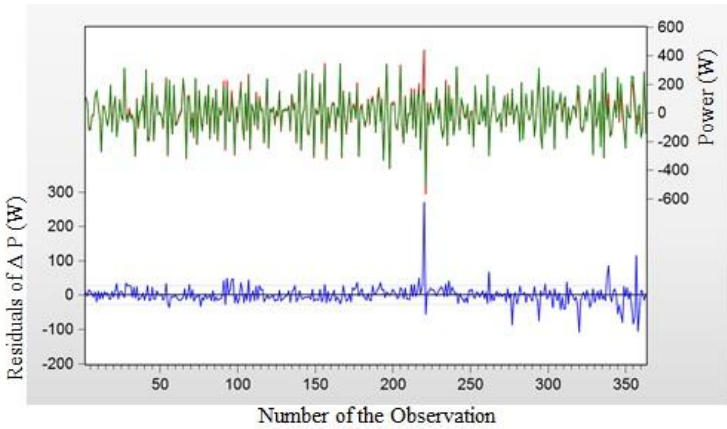

Figure 7. Residual from EG method including first solar irradiance and constant terms.

The colored curves have the same defintion as defined in Figure 3 and where two outliers are identified respectively at position 320 and 357. Thus we applied the EG test for difference regression and the dependent variable is ΔP.

Results are given in Table 16 for the coefficient of the equilibrium error term and the short run term (coefficient of ΔG).

Table 16. Error term regression without constant.

Variable	Coefficient	Std. Error	t-Statistic	Probability
ResidCoint (–1)	–0.601050	0.049178	–12.22193	0.0000
ΔG	0.966657	0.008066	119.8493	0.0000
R-squared	0.977102	Mean dependent var	0.949972	
Adjusted R-squared	0.977037	S.D. dependent var	140.1690	
S.E. of regression	21.24035	Akaike info criterion	8.955253	
Sum squared resid	160610.3	Schwarz criterion	8.976932	
Log likelihood	–1600.990	Hannan-Quinn criter	8.963875	
Durbin-Watson stat	2.062846	F-statistic	7578.573	
Prob (F-statistic)	0.000000			

Dependent Variable: ΔP; Method: Least Squares; Sample: (adjusted 363); Included observations: 362 after adjustments.

The R-squared value is very significant as well as the DW statistic close to 2 there is no more heteroscedasticity as represented in Figure 8.

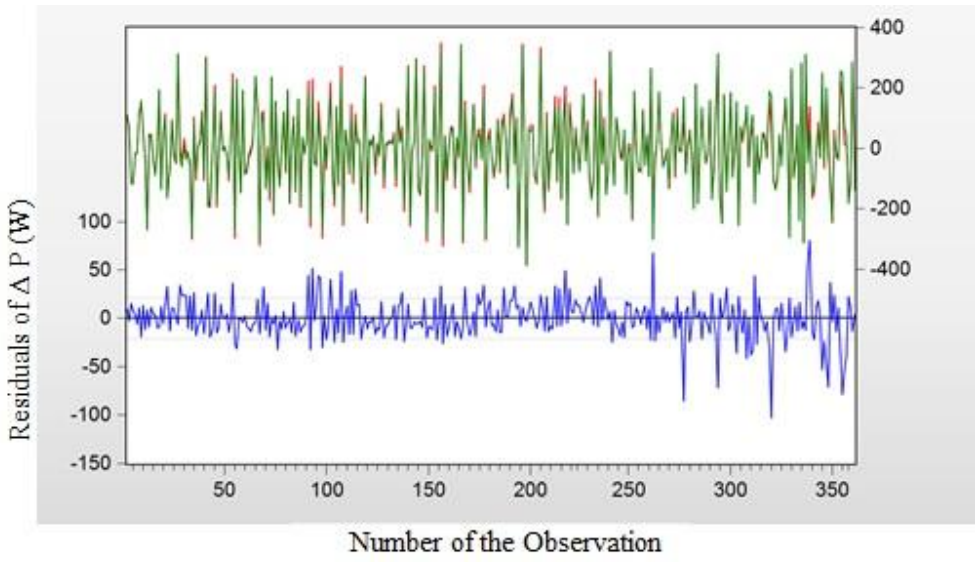

Figure 8. Residual from EG method without suspicious heteroscedasticity.

Thus, the relationship between the variables is given as

$$\Delta P = 0.966\Delta G - 0.601(P_{t-1} - 0.939G_{t-1} - 16.67) \quad (14)$$

where $\Delta P = P_t - P_{t-1}$ and $\Delta G = (G_t - G_{t-1})$ and by substituting ΔP and ΔG in Eq 14 the final relationship for P_t is written as follows

$$P_t = 0{,}966\ G_t - 0.402G_{t-1} + 1{,}601\ P_{t-1} + 10.019 \tag{15}$$

Hence, Eq 15 is the full equation not only with variables P and G in difference but also with P and G with one period lagged.

7. Conclusion and future work

It is difficult to effectively assess the impact of PV output variability on the power grid stability without a clear understanding of the factors that influence this variability. Many factors can become redundant in the presence of other factors and their value in the forecasting framework may vary under different prediction horizons and error measurements. One of the objectives of this paper consisted of applying a statistical method of time series data to identify the most important on-site climatic and environmental parameters that influence PV output variability. Various estimation technique to estimate a linear relationship between PV variables have been proposed in the literature but these OLS regression techniques had lead to spurious results as these techniques were based on the assumption that the variables involved are stationary. Regressions involving time series data include the possibility of obtaining spurious or dubious results in the sense that superficially results look good but on further probing they look suspect.

The originality of this paper is that a technique of cointegration has been introduced known as the EG method according to which models containing nonstationary stochastic variables have been constructed in such a way that the results are statistically meaningful. Through this technique a linear relationship is obtained including not only the dependent variable P(t) and explanatory variable $G_k(t)$, but also their one time lagged variables P(t–1) and $G_k(t-1)$, thus a temporal link between the variables in small samples, that is daily mean data upon one year.

For that, the EG method requires testing whether variables are integrated of the same order. This is done using the ADF unit root test. We then used the error correction model to test whether the residuals of the long run relationship between P and G are stationary. All these conditions are satisfied and the two variables under investigation cointegrate.

The advantage of the Engle-Granger method over the other techniques is its ease of implementation. However, its results are dependent on how the long-run equilibrium equation is specified. In some cases, it might not be easy to identify which variable enters as the dependent variable.

This study was conducted in a temperate region and temperature had been eliminated in the final PV relationship. However, the latter cannot be readily adopted by other countries as both the methodologies employed depend on several site-dependent features. The PV output also is heavily reliant on numerous local meteorological and environmental factors. Therefore, future work has been proposed in a South West Indian Ocean project to improve this PV model.

Acknowledgements

The authors wish to acknowledge the GREEN platform industrial partners.

Conflict of interest

The authors declare no conflict of interest in this paper.

References

1. Agrawal S, Solanki SC, Tiwari GN (2011) Design, fabrication and testing of micro-channel solar cell therrmal (MCSCT) tiles in door condition. World Renewable Energy Congress—Sweden, 2916–2923.
2. Singh GK, Agrawal S, Tiwari A (2012) Analysis of different types of hybrid phtovoltaic thermal air collectors: A comparative study. *J Fund Renew Energ Appl* 2: 1–4.
3. Rajoria CS, Agrawal S, Dash AK, et al. (2015) A newer approach on cash flow diagram to investigate the effect of energy payback time and earned carbon credits on life cycle cost of different photovoltaic thermal array systems. *Sol Energy* 124: 254–267.
4. Ross RGJ (1976) Interface design considerations for terrestrial solar cell modules. Proceedings of the 12th IEEE Photovoltaic Specialists Conference, Baton Rouge LA, USA, 15–18: 801–806.
5. Skoplaki E, Palyvos JA (2009) Operating temperature of photovoltaic modules: A survey of pertinent correlations. *Renew Energ* 34: 23–29.
6. Skoplaki E, Palyvos JA (2009) On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations. *Sol Energy* 83: 614–624.
7. Peled A, Appelbaum J (2017) Enhancing the power output of PV modules by considering the view factor to sky effect and rearranging the interconnections of solar cells. *Prog Photovolt Res Appl* 25: 810–818.
8. Ramenah H, Tanougast C, Cicero L (2014) Toward a prediction of the photovoltaic based power production from Experimental Thermal modeling. Transportation Electrification Asia-Pacific. *IEEE*, 1–4.
9. Skoplaki E, Boudouvis AG, Palyvos JA (2008) A simple correlation for the operating temperature of photovoltaic modules of arbitrary mounting. *Sol Energ Mat Sol C* 92: 139–1402.
10. Jakhrani AQ, Othman AK, Rigitand ARH, et al. (2011) Comparison of solar photovoltaic module temperature models. *World Appl Sci J* 14: 1–8.
11. Radziemska E (2003) The effect of temperature on the power drop in crystalline silicon solar cells. *Renew Energ* 28: 1–12.
12. Singh P, Ravindra NM (2012) Temperature dependence of solar cell performance-an analysis. *Sol Energ Mat Sol C* 101: 36–45.
13. Krauter S, Araújo RG, Schroer S, et al. (1999) Combined photovoltaic and solar thermal systems for facade integration and building insulation. *Sol Energy* 67: 239–248.

14. García MCA, Balenzategui JL (2004) Estimation of photovoltaic module yearly temperature and performance based on nominal operation cell temperatures calculations. *Renew Energ* 29: 1997–2010.
15. Trinuruk P, Sorapipatana C, Chenvidhya D (2009) Estimating operating cell temperature of BIPV modules in Thailand. *Renew Energ* 34: 2515–2523.
16. Savvakis N, Tsoutsos T (2015) Performance assessment of a thin film photovoltaic system under actual Mediterranean climate conditions in the island of Crete. *Energy* 90: 1435–1455.
17. Rosell JI, Ibáñez M (2006) Modelling power output in photovoltaic modules for outdoor operating conditions. *Energ Convers Manage* 47: 2424–2430.
18. Pashiardis S, Kalogirou SA, Pelengaris A (2017) Statistical analysis for the characterization of solar energy utilization and inter-comparison of solar radiation at two sites in Cyprus. *Appl Energ* 190: 1138–1158.
19. Raza MQ, Nadarajah M, Ekanayake C (2016) On recent advances in PV output power forecast. *Sol Energy* 136: 125–144.
20. De Giorgi MG, Congedo PM, Malvoni M (2014) Photovoltaic power forecasting using statistical methods: impact of weather data. *Iet Sci Meas Technol* 8: 90–97.
21. Boland J (2008) Time series modeling of solar radiation. Modeling Solar Radiation at the Earth's Surface. Springer Berlin Heidelberg, 283–312.
22. Steland A (2017) Fusing photovoltaic data for improved confidence intervals. *AIMS Energy* 5: 125–148.
23. Kemmoku Y, Orita S, Nakagawa S, et al. (1999) Daily insolation forecasting using multi-stage neural network. *Sol Energy* 66: 193–199.
24. Mellit A, Menghanem M, Bendekhis M (2005) Artificial neural network model for prediction solar irradiance data: application for sizing stand-alone photovoltaic power system. Power Engineering Society General Meeting. *IEEE* 1: 40–44.
25. Sfetsos A, Coonick AH (2000) Univariate and Multivariate forecasting of hourly solar irradiance with artificial intelligence techniques. *Sol Energy* 68: 169–178.
26. Adel Mellit, Alessandro MP (2010) A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste Italy. *Sol Energy* 84: 807–821.
27. Mahmoud D, Violeta H (2016) Fault detection algorithm for grid connected photovoltaic plants. *Sol Energy* 137: 236–245.
28. Ba M, Ramenah H, Tanougast C (2017) Forseeing energy photovoltaic output determination by a statistical model using real module temperature in the north east of France. *Renew Energ*, in press.
29. Huld T, Amillo AMG (2015) Estimating PV module performance over large geographical regions: The role of irradiance, air temperature, wind speed and solar spectrum. *Energies* 8: 5159–5181.
30. Dickey D, Fuller W (1979) Distribution of the estimates for the autoregressive time series with a unit root. *J Am Stat Assoc* 74: 427–431.
31. Gujarati DN (2004) Basic of Econometric, Fourth Edition. The McGraw-Hill Econometrics, Fourth Companies.

32. Engle RF, Granger CWJ (1987) Co-integration and error correction: representation, estimation, and testing. Essays in econometrics. Harvard University Press, 251–276.
33. Casin P (2009) *Econométrie méthodes et applications avec Eviews*: Editeur Technip. Available from: https://www.eyrolles.com/Entreprise/Livre/econometrie-9782710809272.
34. Ramenah H, Tanougast C, Kalogirou SA, et al. (2016) Reliably model of microwind power energy output under real conditions in France suburban area. *Renew Energ* 91: 1–10.
35. Applied Economic Time Series (1995) Wiley Series in Probability and Statistics: 420.
36. Hamilton JD (1994) Time Series Analysis. Princeton University Press, 767–768.

AIMS Press

© 2018 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)