# Comparison of SARIMAX, SARIMA, Modified SARIMA and ANN-based Models for Short-Term PV Generation Forecasting

Stylianos I. Vagropoulos, G. I. Chouliaras, E. G. Kardakos, C. K. Simoglou, A. G. Bakirtzis

Power Systems Laboratory, Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki
54124, Thessaloniki, GREECE
stelvag@auth.gr, chouligi@gmail.com, ekardako@hotmail.com, chsimoglou@ee.auth.gr, bakiana@eng.auth.gr

***Abstract*- This paper compares four practical methods for electricity generation forecasting of grid-connected Photovoltaic (PV) plants, namely Seasonal Autoregressive Integrated Moving Average (SARIMA) modeling, SARIMAX modeling (SARIMA modeling with exogenous factor), modified SARIMA modeling, as a result of an a posteriori modification of the SARIMA model, and ANN-based modeling. Interesting results regarding the necessity and the advantages of using exogenous factors in a time series model are concluded from this comparison. Finally, intra-day forecasts updates are implemented to evaluate the forecasting errors of the SARIMA and the SARIMAX models. Their comparison highlights differences in accuracy between the two models. All models are compared in terms of the Normalized (with respect to the PV installed capacity) Root Mean Square Error (NRMSE) criterion. Simulation results from the application of the forecasting models in a PV plant in Greece using real-world data are presented.**

***Index Terms*-- Artificial neural networks, autoregressive integrated moving average (ARIMA) modeling, SARIMA modeling, SARIMAX modeling, photovoltaic plants, photovoltaic energy forecasting, solar irradiation.**

## I. INTRODUCTION

The increasing penetration of Renewable Energy Resources (RES) worldwide creates prospects for a cleaner, more sustainable and more decarbonized future. However, apart from the obvious environmental benefits, the large-scale RES penetration in the grid should be carefully addressed, since it brings new challenges in the short-term scheduling and operation of the power systems.

Besides wind parks, photovoltaic (PV) units are continuously gaining acceptance worldwide, since they can be easily installed and operate in various sites [1]. However, although PV generation follows a relative expected daily cycle, it still remains an intermittent and non-dispatchable power source, due to the inherent difficulty in accurately predicting meteorological conditions, such as cloud cover. Therefore, the research community has focused lately on the development of advanced PV forecasting models and tools, which are generally divided into two main groups. The first group comprises the pure physical models, which transform the output of numerical weather prediction (NWP) models into PV power output by performing appropriate post-process calculations and often implementing Model Output Statistics (MOS) to correct the forecasts [2]. A comparison of six different statistical models for very short-term forecasting (0-4 hours ahead) is presented in [3], whereas indicative hybrid forecasting models, that combine different statistical approaches to improve the forecasting efficiency, are described in [4]-[5]. An extensive review of state-of-the art methods for solar irradiation forecasting can be found in [6].

Although the literature concerning the implementation of Seasonal Autoregressive Integrated Moving Average models with exogenous factor (SARIMAX) in PV power generation forecasting is quite limited, there are works which highlight the effectiveness of SARIMAX models in other applications. Indicatively, in [7] a comparison between an Artificial Neural Network (ANN) model and a SARIMAX model is conducted, in order to evaluate their performance in week-ahead forecasting of temperature-driven electricity load. The authors conclude that, although ANN achieves better performance in estimation period, it fails to provide a better forecasting in a week forecasting period compared to the SARIMAX model. In [8], temperature is also used as exogenous factor in a SARIMAX coupled modeling for intraday forecasting of individual load curves. In another application in [9], the authors compare both univariate and multivariate time-series models in daily and monthly snow water equivalent (SWE) forecasting in Ontario, Canada. The results reveal that the SARIMAX model performed better than the SARIMA model.

This paper addresses four practical methods for electricity generation forecasting of grid-connected Photovoltaic (PV) plants and is an extension of our previous work [10]. The first model is based on Seasonal Autoregressive Integrated Moving Average time series analysis (SARIMA modeling). The second model is based on SARIMA time series analysis with exogenous factor (SARIMAX modelling), where short-term solar radiation forecasts derived from Numerical Weather Prediction (NWP) models are used as exogenous inputs. The steps for successfully constructing a SARIMAX model for PV forecasting applications are described in the paper. The third model, the modified SARIMA model, is the result of an a posteriori modification of the SARIMA model, using the available solar radiation forecasts. This model is constructed easier than the pure SARIMAX and the accuracy of the two models is compared. The fourth model adopts

ANN-based models with multiple inputs.

Interesting results regarding the necessity and the advantages of using exogenous factors in a time series model are concluded from this comparison. Finally, intra-day forecasts are implemented to evaluate the forecasting errors of the SARIMA and the SARIMAX models. Their comparison highlights differences in the performance between the two models.

All models are compared in terms of the Normalized (with respect to the PV installed capacity) Root Mean Square Error (NRMSE) criterion. Simulation results from the application of the forecasting models in a PV plant in Greece using real-world data are presented.

## II. PV Forecasting Models

In this section, the main features of the four forecasting models are described in detail.

### A. SARIMA model

A class of time series techniques, namely ARIMA, can be employed for the short-term forecasting of PV power generation. ARIMA is a method first introduced by Box and Jenkins [11] and has now become one of the most popular methods for time series forecasting.

In this paper, a variation of the classical ARIMA model, namely the seasonal ARIMA model (i.e. SARIMA) is used, in order to account for the inherent seasonal effect of the PV power output. The seasonal ARIMA model is generally referred to as $SARIMA\ (p,d,q)\ x\ (P,D,Q)_s$, where p, d, q and P, D, Q are non-negative integers that refer to the polynomial order of the autoregressive (AR), integrated (I), and moving average (MA) parts of the non-seasonal and seasonal components of the model, respectively.

The SARIMA model is described mathematically as follows:

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

where:

$y_t$ is the forecast variable (i.e. PV production)

$\varphi_p(B)$ is the regular AR polynomial of order *p*

$\theta_q(B)$ is the regular MA polynomial of order *q*

$\Phi_P(B^s)$ is the seasonal AR polynomial of order *P*

$\Theta_Q(B^s)$ is the seasonal MA polynomial of order *Q*

The differentiating operator $\nabla^d$ and the seasonal differentiating operator $\nabla_s^D$ eliminate the non-seasonal and seasonal non-stationarity, respectively. B is the backshift operator, which operates on the observation $y_t$ by shifting it one point in time (i.e. $B^k(y_t) = y_{t-k}$). The term $\varepsilon_t$ follows a white noise process and s defines the seasonal period. The polynomials and all operators are defined mathematically as follows:

$$\varphi_p(B) = 1 - \sum_{i=1}^{p}\varphi_i B^i \qquad \Phi_P(B^s) = 1 - \sum_{i=1}^{P}\Phi_i B^{s,i}$$

$$\theta_q(B) = 1 - \sum_{i=1}^{q}\theta_i B^i \qquad \Theta_Q(B^s) = 1 - \sum_{i=1}^{Q}\Theta_i B^{s,i}$$

$$\nabla^d = (1-B)^d \qquad \nabla_s^D = (1-B^s)^D$$

The model development was based on the Box-Jenkins methodology, which consists of four iterative steps: a) Identification, b) Estimation, c) Diagnostic checking and d) Forecasting. Further details on the description of the adopted methodology can be found in [11].

The selection of the most appropriate model was based on the calculation of the day-ahead forecast error using the NRMSE criterion for different model orders and taking into account only the daylight hours. The model with the smallest average yearly NRMSE was selected as the most suitable.

### B. SARIMAX model

SARIMAX model is an extension of the SARIMA model, enhanced with the ability to integrate exogenous (explanatory) variables, in order to increase its forecasting performance. This multivariate version of SARIMA model, called Seasonal ARIMA with eXogenous factor (i.e. SARIMAX), is generally expressed mathematically as:

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D y_t = \beta_k x_{k,t}' + \theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

where $x_{k,t}'$ is the vector including the k[th] explanatory input variables at time *t* and $\beta^k$ is the coefficient value of the k[th] exogenous input variable. The stationarity and invertibility conditions are equal to those of ARMA models.

In the examined application, the response variable of the SARIMAX model is the PV production, while the explanatory variable is the time series of solar radiation forecasts derived from NWP models. As the data is non-stationary (24-hour seasonality included), the key point for the successful construction of SARIMAX model, is the differentiation of both response and exogenous time series before the model's estimation, otherwise there is the hazard of "spurious regression" [12]. The selection process of the seasonal (P, D, Q) and non-seasonal (p, d, q) terms was based mainly in certain information criteria (AIC and FPE), but the autocorrelation (ACF) and partial autocorrelation (PACF) plots were also observed and, therefore, the most suitable candidate models were selected. However, the final selection of the most well-performed SARIMAX model was based on the smallest average yearly NRMSE obtained through a comparison among the aforementioned models.

### C. Modified SARIMA model

In the modified SARIMA modeling, the initial SARIMA model is further improved by incorporating short-term solar radiation forecasts derived from NWP models. For this purpose, the initial hourly forecasts (output of the pure SARIMA model) were multiplied by the following factor:

$$K = \frac{R'_d}{f_1 \cdot R_{d-1} + f_2 \cdot R_{d-2} + .. + f_P \cdot R_{d-P}}$$

where

$R'_d$ denotes the total radiation forecast of day $d$ (forecast day),

$R_{d-i},\ i = 1, 2, ..., P$, denotes the total real radiation of each of the previous $P$ days to day $d$, where $P$ is the order of the seasonal AR polynomial of the SARIMA model, and

$f_i,\ i = 1, 2, ..., P$, are weighting factors calculated on the basis of the respective coefficients of the seasonal AR polynomial of the SARIMA model, as follows:

$$f_i = \Phi_i / \sum_{i=1}^{P} \Phi_i$$

Finally, the modified hourly forecasts are calculated as follows:

$$F'_{SARIMA}(t) = K \cdot F_{SARIMA}(t)$$

### D. Artificial Neural Networks

Artificial Neural Networks (ANNs) have been successfully used in various forecasting applications. They are based on the operation of biological neural networks and supposedly possess the ability of a human-like learning process. A typical ANN structure consists of an input layer, a hidden layer and an output layer. Each layer is comprised of neurons that process the input signals and produce an output, while connections between the layers have a weight factor. An ANN easily adjusts to any set of input-output patterns and through a robust training process forms a model function with the minimum possible error.

Several ANN models were investigated for day-ahead forecasting of power output in single PV parks. The proposed models have one output, namely the forecasted PV power generation P(D,H), where D=forecast day, H=forecast hour.

Thus, the same formed model is used consecutively for all hours of the forecast day. Many inputs were tested based on the autocorrelation function of the power time series and the strong diurnal periodicity. Two of the ANN models are presented in Table I. Both of them have 7 inputs, 1 output and one hidden layer with 4 neurons.

TABLE I
ANN MODELS

| ANN Model | Inputs | Output |
|---|---|---|
| A | P(D-1,H), P(D-1,H-1), P(D-2,H), P(D-3,H), $P_{AVE}(H)$, $R_{MAX}(H)$, R'(D,H) | P(D,H) |
| B | ΔP(D-1,H), ΔP(D-1,H-1), ΔP(D-2,H), ΔP(D-3,H), ΔR'(D,H), ΔR'(D,H-1), ΔR'(D,H+1) | ΔP(D,H) |

where

*P(D,H)* is the power output of the PV plant

$P_{AVE}(H)$ is the average *P(d, H)* for d=D-4,…, D-10

*R'(D,H)* is the available Radiation Forecast

$R_{MAX}(D,H)$ is the maximum (clear sky) radiation for day *D* and hour *H*

*ΔF(D,H)=F(D,H) - F(D-1,H)* is the daily deviation referring either to values of power output (*P*) or radiation forecasts (*R'*).

## III. TEST RESULTS

The developed forecasting models were applied for the day-ahead and intraday hourly PV power generation forecast of one PV plant located in the prefecture of Attica, outside Athens. Its nominal capacity is 0.15 MW, whereas the solar radiation data (real measurements and forecasts) are provided by a weather station located at Spata, very close to the PV plant. The available data set includes: a) the hourly real PV power generation of the PV plant b) the hourly radiation measurements and c) the respective hourly radiation forecasts The available data range covers the time period from 1 Jan 2011 to 31 Dec 2012.

In order to evaluate the performance of the proposed forecasting models, they are all compared to the persistence model. In this paper, persistence is defined as if PV generation in each hour of the forecast day equals the real PV generation of the respective hour of the previous day.

A variety of SARIMA models in terms of the polynomial order of the non-seasonal and seasonal components was examined. The SARIMA $(3,1,2)x(3,1,2)_{24}$ model was found to exhibit the best performance in terms of the average yearly NRMSE.

Regarding SARIMAX modeling, the model, SARIMAX$(5,0,5)x(1,1,1)_{24}$ presented the lowest average yearly NRMSE (11,39%) among the candidate models. This model was then further improved by including two more AR lags, 23 and 47. These two lags were included after the observation of the ACF and PACF plots of the specific model residuals, where some lags showed non-negligible autocorrelation and partial autocorrelation. This indicates that there exists some valuable information not fully captured by the initial model. The inclusion of all those lags was evaluated, however the above two lags resulted in the highest performance improvement. It should also be pointed out that although both AR and MA components were tested, only the inclusion of AR components led to better results. The improved SARIMAX model with the additional AR lags (henceforth "**Improved SARIMAX"**) resulted in a 10,93% NRMSE.

In Figs. 1-3 the performance of the SARIMA and SARIMAX models regarding day-ahead forecasting is compared for three different time periods through the year, i.e. February, April and July, respectively. The forecasts are executed in a rolling manner, starting at midnight (00:00) for 24-hours ahead. It is obvious that the SARIMA model cannot predict accurately the PV power generation when the previous days exhibit irregular production patterns with respect to the forecast day (Fig. 1 and 2), which is dominant mostly in months with changeable weather. On the other hand, SARIMAX model seems to perform much better these days due to the exogenous factor of solar radiation forecasts.

However, this is not the case for consecutive days with similar weather conditions, like in summer. During this season, the SARIMA model outweighs SARIMAX model (Fig. 3). As a consequence, the solar radiation forecast seems not only unnecessary, but its inclusion also decreases the forecasting performance.

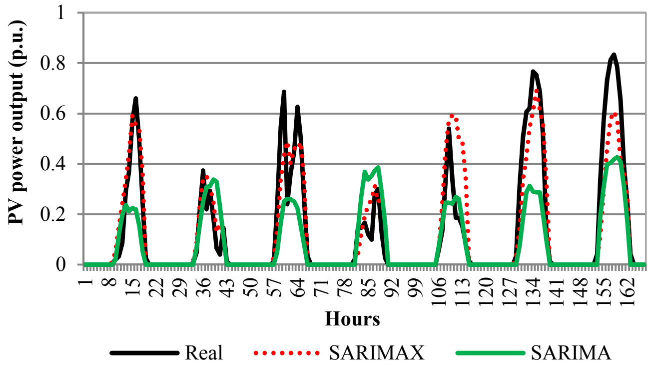


Figure 1. Rolling 24-hours ahead forecasts with SARIMA and SARIMAX models (12/02/2012 – 18/02/2012)

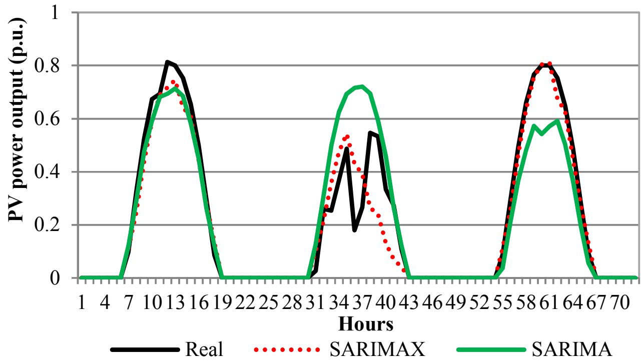


Figure 2. Rolling 24-hours ahead forecasts with SARIMA and SARIMAX models (26/04/2012 – 28/04/2012)

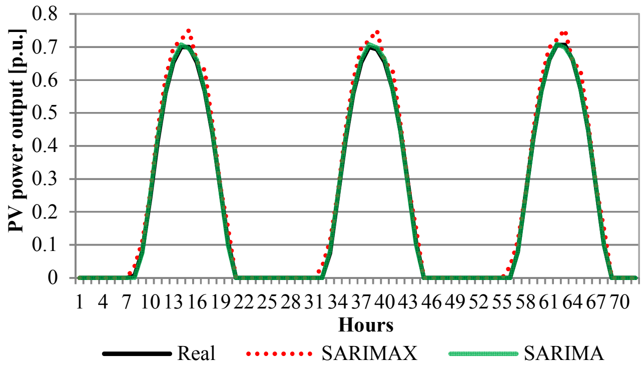


Figure 3. Rolling 24-hours ahead forecasts with SARIMA and SARIMAX models (12/07/2012 – 14/07/2012)

In Figs. 4 and 5 (higher resolution) the performance of the examined forecasting models for a spring day (i.e. 26 May 2012) is examined. Both ANN models, namely A and B, exhibit similar performance throughout the year (see also Table II). For the sake of clarity, only the simulation results obtained with the ANN-Model B are presented in those figures. It is easily concluded that the models that consider the solar radiation forecasts score a better performance for this day. The best performance is scored by the ANN-based model.

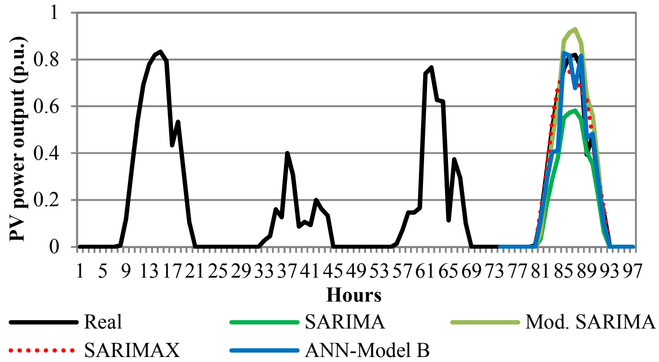


Figure 4. Day-ahead forecasting comparison between the various models - 26/05/2012

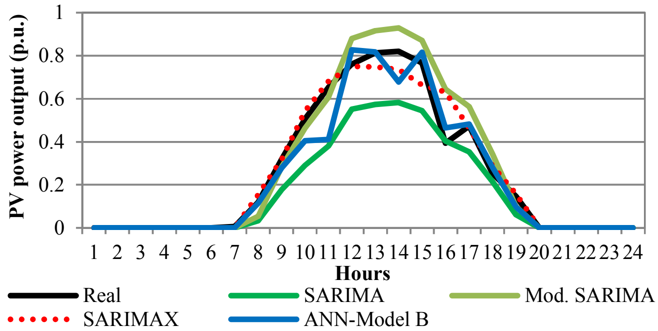


Figure 5. Day-ahead forecasting comparison between the various models - 26/05/2012 (high resolution)

In Table II and Fig. 6 the average seasonal and the average yearly NRMSE for each forecasting model is presented. The model that performs the lowest forecast error in each season is shown in bold. In Fig. 6 the monthly NRMSE is presented.

The aforementioned conclusion that the exogenous variable is unnecessary during months with similar production profiles is further highlighted in both Table II and Fig. 6. In this context, it is obvious that the models which do not include exogenous variable (i.e. Persistence and SARIMA) exhibit a better forecasting performance in summer compared to the other models. On the other hand, models that take into account the radiation forecasts perform much better in winter and spring and slightly better in autumn. A different model scores the best season performance, but the improved SARIMAX model scores the lowest average yearly NRMSE among all models (see Table II).

Another interesting result is the comparison between the improved SARIMAX model and the modified SARIMA model. Both models take into account the radiation forecasts, however the first one endogenously and the second one with an ex-post process after the forecast using the SARIMA model (see Section II.C). The first model performs better for six months and the second performs better for the other six months. Generally, the improved SARIMAX outperforms mostly during winter and spring, whereas the modified SARIMA during summer (see Fig. 6).

The only model which outweighs the performance of the improved SARIMAX model, is a new, optimized combined model, which is created as a combination of two of the aforementioned models, namely the SARIMA model for the months with regular weather patterns (June to September), and the improved SARIMAX model for the rest of the months. The SARIMA model is now trained for optimized performance during the underlying period and not for the whole year. New AR lags are added in the SARIMA model following the procedure previously explained for the improved SARIMAX model.

TABLE II
DAY-AHEAD FORECAST ERRORS

| Forecast Model | | NRMSE [%] | | | | |
|---|---|---|---|---|---|---|
| | | Winter | Spring | Summer | Autumn | Average Yearly |
| Day-Ahead Forecasting | Persistence | 20,35 | 18,00 | **3,17** | 13,33 | 13,71 |
| | SARIMA (3,1,2) x $(3,1,2)_{24}$ | 18,89 | 16,66 | 3,55 | 12,45 | 12,89 |
| | Improved SARIMAX | 14,62 | **11,78** | 4,72 | 12,60 | **10,93** |
| | Modified SARIMA (3,1,2) x $(3,1,2)_{24}$ | 15,06 | 14,02 | 3,61 | **11,82** | 11,12 |
| | ANN - Model A | **14,50** | 12,64 | 5,70 | 12,85 | 11,42 |
| | ANN - Model B | 14,76 | 13,25 | 4,06 | 12,98 | 11,26 |
| | *Optimized combined model (SARIMA and SARIMAX)* | *14,62* | *11,78* | *3,21* | *11,38* | *10,25* |

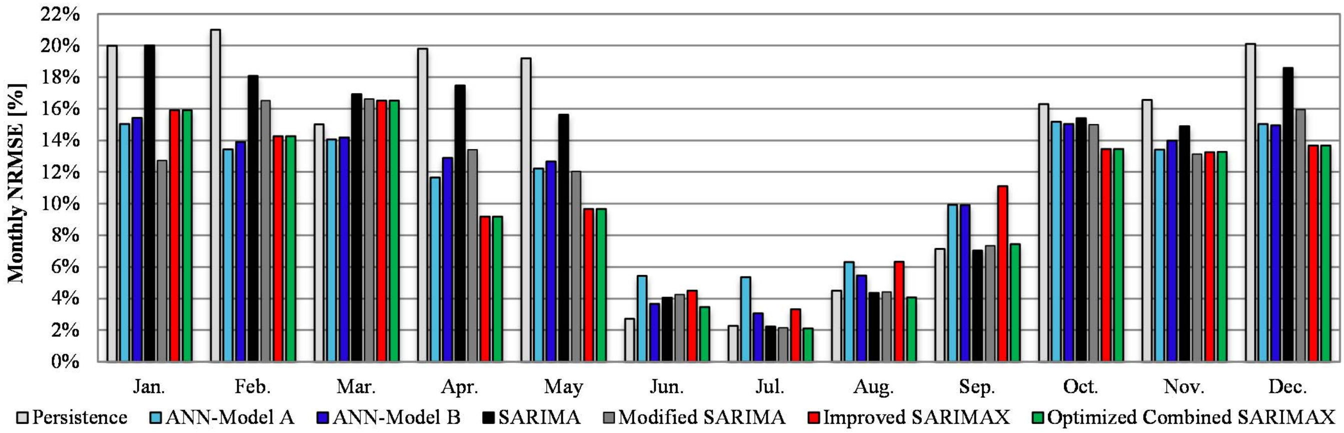


Figure 6. Average monthly NRMSE for all forecasting models

TABLE III
INTRA-DAY FORECAST ERRORS

| Forecast Model | | NRMSE [%] | | | | |
|---|---|---|---|---|---|---|
| | | Winter | Spring | Summer | Autumn | Average Yearly |
| Intra-day forecasting | SARIMA (3,1,2) x $(3,1,2)_{24}$ | **10,52** | 9,85 | **2,95** | **9,14** | **8,12** |
| | Improved SARIMAX | 11,05 | **9,78** | 4,61 | 10,98 | 9,11 |
| | *Optimized combined model* | *9,94* | *9,00* | *2,86* | *8,47* | *7,57* |

In Fig. 7 results from the intra-day forecasting are illustrated and the performance of the SARIMA and the improved SARIMAX model are compared. The forecast is carried out at 12:00 and expands from 13:00 to12:00 of the next day (24-hours ahead). However, for the sake of illustration, in Fig. 7 only the first 12-hours ahead results are presented.

By comparing Fig. 2 and Fig. 7 it is easily conceivable that in the intra-day forecasting case the SARIMA model performs better than in the day-ahead case, since information of the recorded measurements of the first hours of the day (from sunrise to 12:00) are taken into account. The model proved able to adjust to the recent PV generation profile without the necessity of using radiation forecasts. Therefore, the PV power generation forecast for the remaining hours, even for months with highly volatile weather conditions, is improved.

Table III presents the NRMSE calculation for the SARIMA and the SARIMAX models. The annual NRMSE of the SARIMA model is almost 1% smaller than the NRMSE of the SARIMAX case. Therefore, it could be noted that when it comes to intra-day forecasting the exogenous factor could be omitted and the simpler SARIMA model could be successfully used.

An optimized combined model similarly to the day-ahead case was also developed for the case of intra-day forecasting. However, in contrast to the previous case, the SARIMAX model is now used only during April, May and November and the SARIMA model during all the other months, where it performs better. The SARIMAX and the SARIMA models of the intra-day case are the same with the day-ahead case (see Table III).

Finally, it should be noted that all conclusions presented in this paper are based on the data available for the specific site. In addition, the results regarding the value of exogenous variables are based on the utilization of radiation forecasts as exogenous variable. In case of other exogenous variables results might be different.

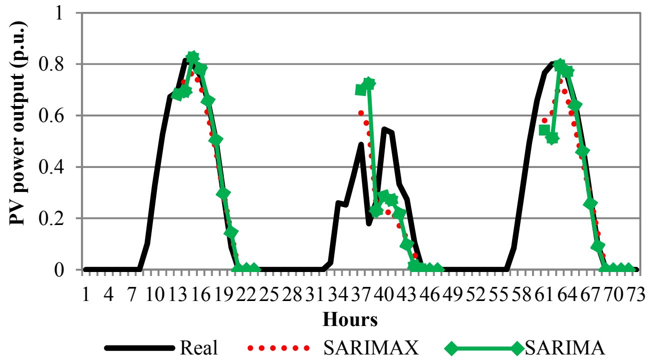

Figure 7. Rolling intra-day forecast, 12-hours ahead starting at 13:00 (26/04/2012 – 28/04/2012)

## IV. Conclusions

This paper presented a fair comparison among four distinct short-term PV generation forecasting models. It is shown that in general, the ANN models, the SARIMAX model and the modified SARIMA model are superior in terms of the day-ahead forecasting performance compared to the persistence model and the SARIMA model. Therefore, the inclusion of solar radiation forecasts improves significantly the day-ahead forecasting accuracy of the PV power generation. The only exception lies in summer, when the persistence model and the SARIMA model exhibit the lowest forecasting NRMSE and, therefore, can successfully be utilized during days with clear sky conditions, which prevail during summer in Greece. However, their performance during the other seasons degrades noticeably. Finally, regarding the intra-day forecasting, the univariate SARIMA model performs better than the multivariate SARIMAX model for the majority of the months and, therefore, the exogenous variable could be omitted.


## Acknowledgements

The authors are thankful to the Independent Power Transmission Operator of Greece (IPTO or ADMIE) S.A. and the National Observatory of Athens for providing the PV production data and solar radiation data, respectively.


## References


[1] IEA-PVPS, "2014 Snapshot of Global PV Markets," Report IEA PVPS T1-26:2015, Mar. 2015. [Online]. Available: http://www.iea-pvps.org/fileadmin/dam/public/report/technical/PVPS_report_-_A_Snapshot_of_Global_PV_-_1992-2014.pdf

[2] P. Mathiesen and J. Kleissl, "Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States," *Solar Energy*, vol. 85, issue 5, pp. 967-977, May 2011.

[3] G. Reikard, "Predicting solar radiation at high resolutions: A comparison of time series forecasts," *Solar Energy*, vol. 83, issue. 3, pp. 342-349, Mar. 2009.

[4] A. Mellit, M. Benghanem, A. H. Arab, and A. Guessoum, "A simplified model for generating sequences of global solar radiation data for isolated sites: Using artificial neural network and a library of Markov transition matrices approach," *Solar Energy*, vol. 79, no. 5, pp. 469-482, Nov. 2005.

[5] M. Cococcioni, E. D'Andrea, and B. Lazzerini, "24-hour-ahead forecasting of energy production in solar PV systems," in *Proc. of the International Conference on Intelligent Systems Design and Applications 2011*, Cordoba, Spain, Nov. 22–24, 2011.

[6] H.M. Diagne, M. David, P. Lauret, and J. Bolan, "Solar irradiation forecasting: State-of-the-art and proposition for future developments for small-scale insular grids," in *Proc. of WREF 2012*, Denver, Colorado, May 2012.

[7] N. Liu, V. Babushkin and A. Afshin, "Short-Term Forecasting of Temperature Driven Electricity Load Using Time Series and Neural Network Model", *Journal of Clean Energy Technologies,* vol. 2, no.4, Oct. 2014.

[8] S. Bercu, and F. Proïa, "A SARIMAX coupled modelling applied to individual load curves intraday forecasting", *Journal of Applied Statistics*, Jul. 2014

[9] R. Kelly, R. Modarres and Ali Sarhadi, "Snow water equivalent time-series forecasting in Ontario, Canada, in link to large atmospheric circulations", *Hydrological Processes,* vol. 28, no. 16, pp. 4640-4653, Aug. 2014.

[10] E. G Kardakos, M. C. Alexiadis, S. I. Vagropoulos, C. K. Simoglou, P. N. Biskas, A. G. Bakirtzis, "Application of time series and artificial neural network models in short-term forecasting of PV power generation," Power Engineering Conference (UPEC), 2013 48th International Universities', Dublin, Sep. 2013.

[11] G.E. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*, Holden Day, 1976.

[12] Rob J. Hyndman. (2010, Oct.), *The ARIMAX Model Muddle,* [Online]. Available: http://robjhyndman.com/hyndsight/arimax/