

# Assessment of forecasting techniques for solar power production with no exogenous inputs

Hugo T.C. Pedro, Carlos F.M. Coimbra \*

*Department of Mechanical and Aerospace Engineering, Jacobs School of Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0411, USA*

Received 2 October 2011; received in revised form 14 March 2012; accepted 6 April 2012  
Available online 4 May 2012

Communicated by: Associate Editor David Renne

## Abstract

We evaluate and compare several forecasting techniques using no exogenous inputs for predicting the solar power output of a 1 MWp, single-axis tracking, photovoltaic power plant operating in Merced, California. The production data used in this work corresponds to hourly averaged power collected from November 2009 to August 2011. Data prior to January 2011 is used to train the several forecasting models for the 1 and 2 h-ahead hourly averaged power output. The methods studied in this work are: Persistent model, Auto-Regressive Integrated Moving Average (ARIMA), k-Nearest-Neighbors (kNNs), Artificial Neural Networks (ANNs), and ANNs optimized by Genetic Algorithms (GAs/ANN). The accuracy of the models is determined by computing error statistics such as mean absolute error (MAE), mean bias error (MBE), and the coefficient of correlation ( $R^2$ ) for the differences between the forecasted values and the measured values for the period from January to August of 2011. This work also addresses the accuracy of the different methods as a function of the variability of the power output, which depends strongly on seasonal conditions. The findings show that the ANN-based forecasting models perform better than the other forecasting techniques, that substantial improvements can be achieved with a GA optimization of the ANN parameters, and that the accuracy of all models depends strongly on seasonal characteristics of solar variability.

© 2012 Elsevier Ltd. All rights reserved.

**Keywords:** Solar forecasting; Solar energy; Regression analysis; Stochastic learning

## 1. Introduction

One of the critical challenges in transitioning to an energy economy based on renewable resources is to overcome issues of variability, capacity and reliability of non-dispatchable energy resources such as solar, wind or tidal. The variable, and sometimes intermittent, nature of these resources implies substantial challenges for the current modus operandi of power producers, utility companies and independent service operators (ISOs); especially when high market penetration rates (such as the ones now

mandated by law in California and other US states) are considered. As of January of 2012, the US Department of Energy lists 29 states with Renewable Portfolio Standards (RPSs) varying from 10% to 33% renewable penetration for 2020 to 40% in 2030. Other US states have similar goals. Many European countries have more aggressive RPS goals. Most of the future renewable generation needed for satisfying these aggressive RPS goals will likely come from variable resources such as solar and wind power.

Although solar energy is clearly the most abundant power resource available to modern societies, the implementation of widespread solar power utilization is so far impeded by its sensitivity to local weather conditions, intra-hour variability, and dawn and dusk ramping rates.

\* Corresponding author.

E-mail address: [ccoimbra@ucsd.edu](mailto:ccoimbra@ucsd.edu) (C.F.M. Coimbra).

The variability directly affects both capital and operational costs, also contributing to lower capacity factors. Solar forecasting, i.e., the ability to forecast the amount of power produced by solar farms and rooftop installation feeding power substations, has the ability to optimize decision making at the Independent System Operator (ISO) level by allowing corrections to unit commitments and out-of-region trade. Short-term, intra-hour forecasts are relevant for dispatching, regulatory and load following purposes, but the intra-day (especially in 1–6 h ahead horizon) forecasts are critical for power system operators that handle multiple load zones, and trade outside of their territory. In particular, the direct sunlight beam, which is critical for concentrating solar technologies, is much less predictable than the global irradiance, which includes the diffuse component from the sky hemisphere (see e.g., [Marquez and Coimbra, 2011](#)), and is also more susceptible to truly intermittent availability due to cloud cover. So in order to characterize the viability of solar production in a given location, detailed variability and forecastability studies become imperative. [Fig. 1](#) depicts the variations of 100 kW to 600 kW (i.e. variations from 10% to 60% of the nominal PV plant peak output) during the diurnal period of 9:00–14:00 in a monthly basis. As expected for California's Central Valley, larger fluctuations occur in the Winter, late Fall and early Spring. The Months of July and August show much smaller variability. However, even during some periods within the sunshine months, sudden changes on the power output can be observed as exemplified by an event in August 28, 2010, in which there was an integrated 60% ramp between 10:00 am and 11:00 am (this event is depicted in [Fig. 2](#)).

To understand and predict the variability of the solar resource, many attempts to forecast solar irradiance (the resource) have been presented ([Mellit, 2008](#); [Mellit and Pavan, 2010](#); [Marquez and Coimbra, 2011](#); [Elizondo et al., 1994](#); [Mohandes et al., 1998](#); [Hammer et al., 1999](#); [Sfetsos and Coonick, 2000](#); [Paoli et al., 2010](#); [Lara-Fanego et al., 2011](#)), while other researchers have extended their models to power output from PV plants ([Picault et al., 2010](#); [Bacher et al., 2009](#); [Chen et al., 2011](#); [Chow et al.,](#)

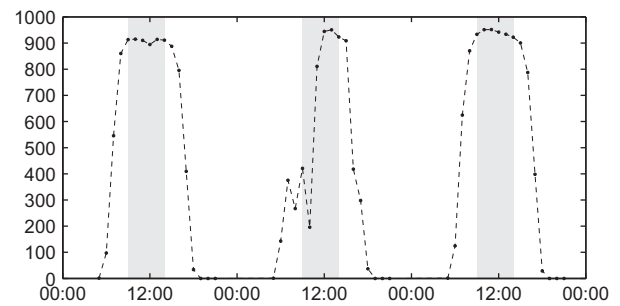


Fig. 2. Power output for the period between 08/27/2010 and 08/29/2010. August 28 exemplifies a day in which there is a sudden ramp up in the power output of more than 600 kW (60% of the nominal peak  $P$ ) within the period of expected maximum Power output.

[2011](#); [Martín et al., 2010](#)). Artificial Neural Networks (ANNs), Fuzzy Logic (FL) and hybrid systems (GA/ANN, ANN-FL) are well suited to model the stochastic nature of the underlying physical processes that determine solar irradiance at the ground level (and thus the power output of PV installations). Other regression methods often employed to describe complex non-linear atmospheric phenomena include the Auto-Regressive Moving Averages (ARMAs) method, as well as its variations, such as the Auto-Regressive Integrated Moving Averages (ARIMAs) method ([Gordon, 2009](#)).

In this work, the 1-h averaged data for the 1 MWp PV farm power output ( $P$ ) collected from November 2009 to August 2011 is used to develop and train several forecasting models for predicting the power output 1 and 2 h-ahead of time, using only the single-axis panels as network of ground sensors. The goal is to study several of the most popular forecasting methodologies and assess their accuracy in order to determine a minimum performance level for which comparison with more complex forecasting methods (using a variety of radiometric and meteorological inputs) should be carried out. In essence, we apply different forecasting engines to a “zero-telemetry” operational data set to evaluate the performance of non-exogenous methodologies, thus creating a baseline for developing more sophisticated.

## 2. Data

This work uses data collected from a single-axis tracking, polycrystalline photovoltaic, 1 MW peak solar power plant located in Central California (Merced). This solar farm provides between 17% and 20% of the power consumed yearly by the University of California, Merced campus, and is used as a test-bed for solar forecasting and fast demand response studies by our research group. The time period analyzed spans from November 3, 2009 (the first full day of operation of the solar farm) to August 15, 2011. The data points collected from the power plant site correspond to the hourly average of power output ( $P$ ). Although available to us, additional solar irradiance and weather variables, such as global horizontal irradiance, cloud cover,

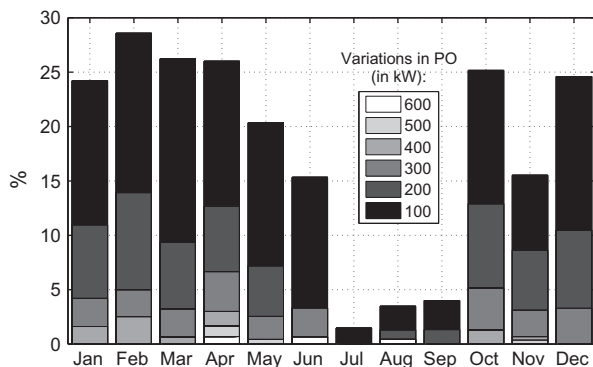


Fig. 1. Percentage of variations in the  $P$  of more than: 100 kW, 200 kW, ..., 600 kW, in a monthly basis. Only data in the peak production period between 09:00 and 14:00 was used to generate this figure.

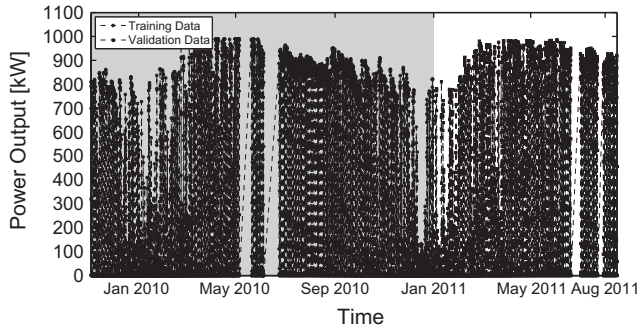


Fig. 3. Hourly averaged power output ( $P$ ) from November 2009 to August 2011. The data in the shaded area was used to create the several forecasting models the remainder was used to validate the models. The gaps in the figure represent malfunction or maintenance periods for the power plant.

wind speed and direction are not considered in this study because the objective is to assess performance of various univariate, endogenous methodologies in the forecasting of the power output.

The power output data set is plotted in Fig. 3. The data points in the shaded area are used to develop the various forecasting models below (e.g. to train the ANNs or to find the ARIMA coefficients), and the remainder are used to assess the performance of the methods when compared with measured data through various statistical metrics. The gaps visible in Fig. 3 correspond to periods in which the power plant was not operating at full capacity due to malfunction or maintenance.

### 3. Methodology

Five different methods to forecast the power output of the 1 MWp PV power plant, 1 and 2 h ahead of time are used in this work. The methods employed are:

- Persistent model.
- Autoregressive Integrated Moving Average (ARIMA).
- k-Nearest-Neighbors (kNNs).
- Artificial Neural Networks (ANNs).
- and an hybrid Genetic Algorithm/Artificial Neural Network (GA/ANN).

Given that the stochastic process which generates the solar power is not stationary as it is a function of the daily solar power we apply a clear sky decomposition to the power output data before applying the listed forecasting methodologies. In the following sections we explain the calculation of the clear sky model and the details of the different forecasting models.

#### 3.1. Clear-sky model

The power output of the PV power plant is a function of the location, the time, the PV technology used, the area of the panels, their orientation and current atmosphere

conditions. In principle the dependence of the power output with respect to all these variables with the exception of the current sky conditions can be modeled deterministically. By assuming clear-sky conditions the power output no longer depends upon this stochastic variable and the resulting model is designated as the Clear-Sky model for the power output. An explicit, analytical expression for the clear-sky model would require detailed knowledge of all the deterministic variables that we do not possess, therefore we resort to an approximated function for the clear-sky model.

The first step to obtain the model is to plot the  $P$  time series from Fig. 3 as a function of the time of the day  $\tau_D$  (as a fraction of the whole day with 0 as the beginning of the day and 1 as the end of the day) and the day of the year  $\tau_Y$ . Both variables ( $\tau_D, \tau_Y$ ) can be easily calculated from the variable  $t$  (where  $t$  is given as a serial date number format):

$$\tau_D = t - \lfloor t \rfloor \quad (1)$$

$$\tau_Y = \lfloor t \rfloor - t_{Y-01-01} \quad (2)$$

where  $t_{Y-01-01}$  represents the serial number for the first day of the year  $Y$ . In case there are multiple values of  $P$  for a given pair of ( $\tau_D, \tau_Y$ ) the plotted value of  $P$  corresponds to their average. The output of this operation is depicted in Fig. 4 (top).

In the second step, a smooth surface that envelops closely the measured power output shown in Fig. 4 (top) is created manually. This surface is shown in Fig. 4 (bottom) and it corresponds to the clear sky model  $P_{c-s}(\tau_D, \tau_Y)$ . No analytical expression was obtained for this function, instead we interpolate linearly the points depicted in

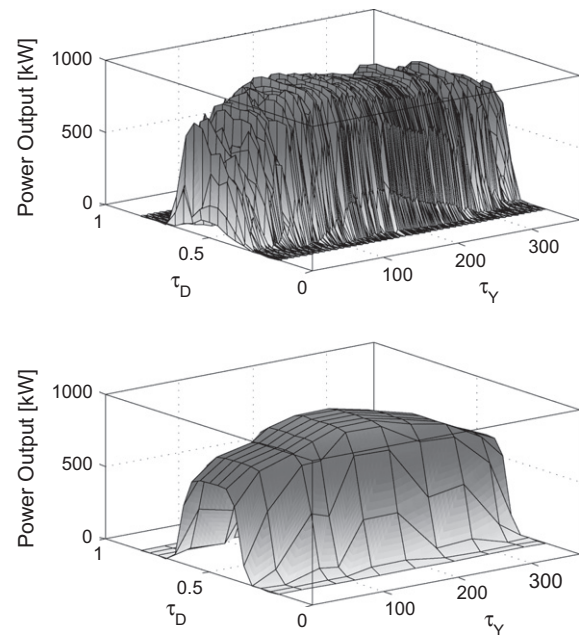


Fig. 4. Top: the measured power output as a function of the time of the day  $\tau_D$  and the day of the year  $\tau_Y$ . Bottom: the power output expected under clear sky conditions as a function of the same variable.

Fig. 4 (bottom) whenever it's necessary to calculate  $P_{c-s}(\tau_D, \tau_Y)$ .

After computing the clear-sky power output model the original  $P$  time series can be decomposed as

$$P(t) = P_{c-s}(t) + P_{st}(t) \quad (3)$$

where the stochastic difference to the clear-sky model denoted by  $P_{st}$ . All the forecasting models with the exception of the persistent model will operate with the stochastic component of  $P$ .

### 3.2. Persistent model

One of the simplest models for the forecasting of a time series is the so called Persistent model. As the name indicates in the Persistent model the future values of the time series are calculated assuming that conditions remain unchanged between time  $t$  and the future time  $t + \Delta t$  (where  $\Delta t$  can be 1 or 2 h in this work). If the time series under study is stationary the most logical implementation of the persistent model is,  $\hat{P}(t + \Delta t) = P(t)$  (where  $\hat{\cdot}$  denotes a forecasted variable). However in the current case the power output time series is not stationary and a better implementation of the persistent model is:

$$\hat{P}(t + \Delta t) = \begin{cases} P_{c-s}(t + \Delta t), & \text{if } P_{c-s}(t) = 0 \\ P_{c-s}(t + \Delta t) \frac{P(t)}{P_{c-s}(t)}, & \text{otherwise} \end{cases} \quad (4)$$

where  $P_{c-s}(t)$  is the expected power output under clear-sky conditions as described in the previous section. This model amounts to say that the fraction of the power output relative to the clear-sky conditions remains the same between times  $t$  and  $t + \Delta t$ . In case the current  $P_{c-s}(t)$  is zero (at night) the forecasted value is taken as the value for clear sky conditions.

### 3.3. Autoregressive integrated moving average

For non-stationary time series one of the most popular statistical forecasting tools is a class of ARIMA models (Brockwell and Davis, 2002; Box et al., 2008). ARIMA models couple an autoregressive component (AR) to a moving average component (MA). Fig. 5 shows the correlation

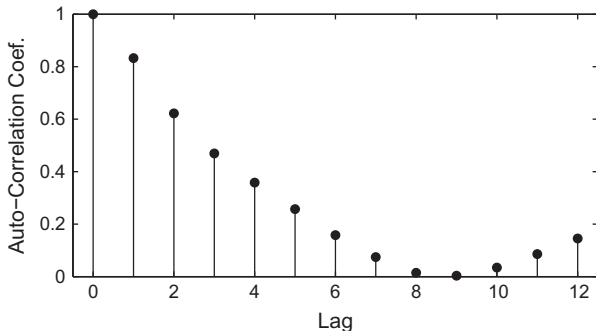


Fig. 5. Autocorrelation plot for the entire the stochastic component  $P_{st}$  for 1 h lags.

coefficient for several time lags of the stochastic variable  $P_{st}$ . The autocorrelation plot shows that the sample autocorrelations are strong and decay slowly and it indicates that the process is non-stationary validating the application of the ARIMA model to this problem. The ARIMA model with parameters  $p$ ,  $d$  and  $q$  can be written as:

$$Y_t = (1 - B)^d X_t \quad (5)$$

$$Y_t = \sum_{i=1}^p \phi_i Y_{t-i} + \sum_{j=1}^q \theta_j u_{t-j} \quad (6)$$

where  $B$  is the backward operator (e.g.  $B(Y_t) = (Y_t - Y_{t-1})$ ),  $u_t$  is an error term distributed as a Gaussian white noise, and the parameters  $p$ ,  $d$  and  $q$  and the coefficients  $\phi_i$  and  $\theta_j$  can be determined using various model identification tools (Box et al., 2008). Here we use the open source code *R* (version 2.13.1) which includes an implementation of the ARIMA method. *R* was used together with Statconn, which enables to execute R commands from within MATLAB. This facilitated the manipulations of large data sets and the processing of results. Hundreds of combinations of the parameters ( $p, d, q$ ) were tested and the best model was chosen based on the Mean Square Error (MSE) criterion between the ARIMA forecasting and the training data. Once the model was established the forecasting for the stochastic component of the power output  $\hat{P}_{st}$  for the validation data was obtained by solving (5) and (6) with  $X_t = \hat{P}_{st,t}$ .

### 3.4. k-Nearest-neighbors

The k-Nearest-Neighbors algorithm (kNN) is one of the simplest methods among the machine learning algorithms. It is a pattern recognition method for classifying patterns or features (Duda and Hart, 2000). The classification is based on the similarity of the pattern to classify with respect to training samples in the feature space. For this problem the kNN model consists of looking into the history of the  $P_{st}$  time-series and identifying the timestamp that resembles the “current” conditions most closely. The forecasted value (which corresponds to the classification of the pattern) is taken as the subsequent value in the time-series. The best match is found by comparing the current vector of features or patterns ( $\vec{Q}_t$ ) against the historical patterns. In this work  $\vec{Q}_t$  is defined as

$$\vec{Q}_t = (P_{st,t}, P_{st,t-1}, \dots, P_{st,t-N}), \quad (7)$$

where  $N$  is 12. And the forecasting is

$$\hat{P}_{st,t+1} = P_{st,K+1} \quad (8)$$

where  $K$  is such that

$$\sqrt{\sum_i (Q_{t,i} - Q_{K,i})^2} \leq \sqrt{\sum_i (Q_{t,i} - Q_{k,i})^2}, \quad k = 1, \dots, n; \quad k \neq K \quad (9)$$

where the features space  $\vec{Q}_k$  contains only training data - data from before January 2011, and  $n$  is the size of the training data set. In case Eq. (9) returns more than one



value the forecasted value is taken as the average of the various  $P_{st,K+1}$ . The kNN method was also implemented using MATLAB.

### 3.5. Artificial neural network

Artificial neural networks (Bishop, 1995) are useful tools for problems in classification and regression, and have been successfully employed in forecasting problems Marquez and Coimbra (2011) and Mellit and Pavan (2010). Extensive reviews on the forecasting with ANNs can be found in Zhang et al. (1998) and Mellit (2008) with the latter focusing exclusively on solar radiation modeling.

One of the advantages of ANNs is that no assumptions are necessary about the underlying process that relates input and output variables. In general, neural networks map the input variables to the output by sending signals through elements called neurons. Neurons are arranged in layers, where the first layer receives the input variables, the last produces the output and the layers in between, referred to as hidden layers, contain the hidden neurons. A neuron receives the weighted sum of the inputs and produces the output by applying the activation function to the weighted sum. Inputs to a neuron could be from external stimuli or could be from output of the other neurons. Once the ANN structure, number of layers, number of neurons, activation functions and so forth are established the ANN undergoes a training process in which the weights that control the activation of neurons are adjusted so that the minimization of some performance function is achieved, typically the mean square error (MSE). Numerical optimization algorithms such as back-propagation, conjugate gradients, quasi-Newton, and Levenberg–Marquardt have been developed to effectively adjust the weights.

Generally, the ANN forecasting model for the stochastic component of the power output  $P_{st}$  based on can be written as

$$\hat{P}_{st,t+1} = f(\vec{X}). \quad (10)$$

where  $\vec{X} = (P_{st,t}, P_{st,t-1}, \dots, P_{st,t-N})$ , and  $N = 12$  in this work. This work uses several time lagged values of  $P$  in order to take advantage of the autoregressive nature of the time series. The actual form for  $f$  depends on the ANN architecture and transfer functions.

The performance of the ANN depends strongly on its structure as well as the choice of activation functions and, especially, the training method. For this study the following settings for the ANN were selected based on the authors previous experience:

- the ANN is a feed-forward network with 1 hidden layer with 20 neurons, 1 output neuron ( $\hat{P}_{t+1}$ ) and 13 input neurons ( $\vec{X}$ ).
- The activation function for the hidden layer is the hyperbolic tangent sigmoid transfer function and the activation function for the output layer is the linear transfer function.

- The ANN is trained with the Levenberg–Marquardt backpropagation algorithm based on the *MSE* performance.
- Eighty percent of the historical data is used to train the ANN, the remainder 20% are used for testing.

The ANN model was implemented in MATLAB using the Neural Network toolbox 6.0.

In addition to the factors pointed above the performance of the ANNs depend strongly on the input variables. There are several tools (for example normalization, principal component analysis Bishop (1995) and the Gamma test for input selection Marquez and Coimbra (2011)) to pre-process the input data to increase the forecasting performance. In this work only input normalization is applied to the input data. Normalization maps all the input variables into the interval  $[-1, 1]$  following a linear transformation.

### 3.6. Genetic algorithm/artificial neural network

As highlighted above the usage of ANN requires the user to make several decisions with little justification that may lead to a suboptimal model. One way to avoid a trial and error procedure and possibly obtain a big increase in the performance of the ANN is to couple the ANN with some optimization algorithm such as Genetic Algorithms (GAs) (Castillo et al., 2000; Armano et al., 2005).

Genetic algorithms are biological metaphors that combine an artificial survival of the fittest with genetic operators abstracted from nature Holland (1975). In this solution space search technique, the evolution starts with a population of individuals, each of which carries a genotypic and a phenotypic content. The genotype encodes the primitive parameters that determine an individual layout in the population. In this work the genotype consist of weights  $w_j \in \{0, 1\}$ ,  $j = 1, \dots, N + 1$  that control the inclusion/exclusion of the various time lagged values of  $P_t$  in the list of ANN input variables, the number of hidden layers and the number of neurons in each layer which may assume values in the ranges  $[1, 4]$  and  $[4, 30]$ , respectively.

For this case the mathematical formulation of the forecasting model becomes

$$\hat{P}_{st,t+1} = g(\vec{Y}), \quad (11)$$

where the input vector is a subset of the input vector in Eq. (10) determined by the GA variable  $w$  as  $\vec{Y} = X_i|_{w_i=1}$  for  $i = 1, \dots, N + 1$ , and the functional form of  $g$  depends upon the number of hidden layers and the number of neurons in each hidden layer.

For the purpose of the GA, each individual in the population is represented by a vector of integers with  $N + 3$  elements. The first  $N + 1$  correspond to the binary weights  $w$  and the last two elements correspond to the number of hidden layers and the number of neurons per layer, respectively. With these parameters a ANN is created and a fore-

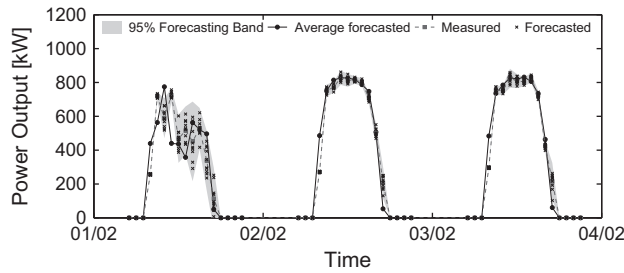


Fig. 6. Time series plot from the best ANN obtained with the GA for the 1 h ahead forecasting. The ANN was trained 10 times and the outputs of each training are shown with the symbols (x). The 95% forecasting error band was computed for each timestamp.

casting model is produced. The MSE between the measured values and the forecasted values is used as the fitness of the GA individual. The GA optimizes these parameters by evolving an initial population based on the selection, crossover and mutation operators with the objective to minimize this forecasting error.

### 3.6.1. Selection, crossover, mutation and stopping criterion

An initial population of 50 individuals is generated randomly with a uniform distribution. Crossover operates on individuals (parents) determined by the selection operator. Selection discovers the good features in the populations based on the fitness value of the individuals. The selection method used here is the tournament method, in which groups of four individuals are randomly selected to play a “tournament”, where the best fit is selected. The fitness is evaluated by calculating the MSE between the forecasted data and training data. The tournaments continue until a predetermined percentage of the population is selected as parents for crossover.

Crossover then proceeds to recombine the “genetic material” of the selected parents. This work uses the scattered

method in which a random binary vector with the same length of the genome is used to select the genes coming from each parent. The crossover operator selects genes from the first parent where the vector has 0 entries and selects genes from the second parent when it has one entries.

Mutation operates on the individuals that have not been selected for reproduction. In this work mutation is achieved by adding a random number with a Gaussian distribution to every gene in the genome. The Gaussian distribution has zero mean and a standard deviation that shrinks as the number of generations increases. In this work the ratio crossover/mutation is 4/1.

Once the population for a new generation is determined this process continues until some criterion is met. Because it is usually difficult to formally specify a convergence criterion for the genetic algorithm because of its stochastic nature, in this work the algorithm stops after 50 generations or if no improvement has been observed over a pre-specified number of generations, in this case 20, whichever is encountered first.

## 4. Results and discussion

As described above all the forecasting models with the exemption of the persistent model forecast the stochastic component of the power output  $P_{st}$ , however the error analysis and the results will be presented in terms of the actual power output which can be easily obtained by  $\hat{P}(t) = P_{c-s}(t) + \hat{P}_{st}(t)$ .

### 4.1. ANN models

The ANN training with the Levenberg–Marquardt algorithm requires the random initialization of the ANN’s weights and biases. This prevents results from being exactly reproducible given that the algorithm may converge to dif-

Table 1  
Parameters for the best ANN obtained with the genetic algorithm.

Forc. horizon	Input Selection													ANN architecture	
	$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	$w_6$	$w_7$	$w_8$	$w_9$	$w_{10}$	$w_{11}$	$w_{12}$	$w_{13}$	No. of layers	No. of neurons/layer
1 h	1	1	1	1	1	1	1	1	0	1	0	1	0	3	12
2 h	1	1	1	0	1	0	0	1	0	0	0	0	1	3	16

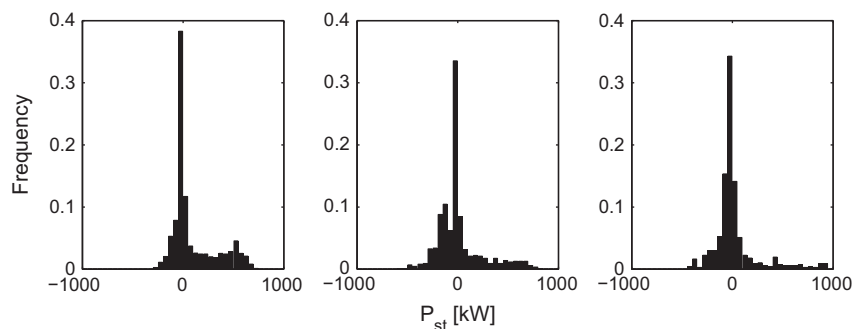


Fig. 7. Probability density function for the stochastic component  $P_{st}$  for the three variability periods: left “ $P_1$ ”, middle “ $P_2$ ” and right “ $P_3$ ”.

Table 2

Statistical error metrics for the 1 h ahead forecasting for the several methodologies. The highlighted numbers identify the best performing method for a given error metric and variability period.

Model	MAE (kW)				MBE (kW)				RMSE (kW)				nRMSE (%)				$R^2$			
	Tot	$P_1$	$P_2$	$P_3$	Tot	$P_1$	$P_2$	$P_3$	Tot	$P_1$	$P_2$	$P_3$	Tot	$P_1$	$P_2$	$P_3$	Tot	$P_1$	$P_2$	$P_3$
Persistent	61.65	61.28	66.90	56.08	29.46	24.46	32.51	40.84	107.48	109.83	110.06	96.31	19.27	22.80	17.35	14.60	0.92	0.91	0.92	0.94
ARIMA	72.80	79.60	73.00	51.78	<b>−0.50</b>	−0.92	<b>−0.52</b>	<b>0.80</b>	105.68	115.57	104.24	69.77	18.95	23.99	16.43	10.58	0.92	0.90	0.93	0.97
kNN	61.92	71.65	69.15	<b>22.94</b>	−0.55	2.38	−4.51	−4.45	116.54	129.18	124.09	<b>42.07</b>	20.90	26.82	19.56	<b>6.38</b>	0.91	0.87	0.90	<b>0.99</b>
ANN	53.49	61.23	53.76	29.51	1.60	−1.61	0.34	13.01	88.23	98.22	87.63	47.21	15.82	20.39	13.81	7.16	0.95	0.93	0.95	0.98
GA/ANN	<b>42.96</b>	<b>48.90</b>	<b>42.98</b>	24.76	1.08	<b>0.52</b>	−2.08	6.85	<b>72.86</b>	<b>80.56</b>	<b>72.45</b>	42.19	<b>13.07</b>	<b>16.72</b>	<b>11.42</b>	6.40	<b>0.96</b>	<b>0.95</b>	<b>0.97</b>	<b>0.99</b>

Table 3

Statistical error metrics for the 2 h ahead forecasting for the several methodologies. The highlighted numbers identify the best performing method for a given error metric and variability period.

Model	MAE (kW)				MBE (kW)				RMSE (kW)				nRMSE (%)				$R^2$			
	Tot	$P_1$	$P_2$	$P_3$	Tot	$P_1$	$P_2$	$P_3$	Tot	$P_1$	$P_2$	$P_3$	Tot	$P_1$	$P_2$	$P_3$	Tot	$P_1$	$P_2$	$P_3$
Persistent	91.12	91.74	95.32	83.88	44.19	37.84	45.47	61.92	160.79	164.33	160.93	149.29	28.86	34.11	25.36	22.70	0.83	0.79	0.83	0.85
ARIMA	102.76	113.83	102.76	68.96	−0.65	−1.90	<b>−0.11</b>	<b>2.48</b>	144.26	157.97	142.70	93.44	25.89	32.79	22.49	14.21	0.86	0.81	0.87	0.94
kNN	87.76	104.41	92.71	30.58	−3.44	−0.81	−8.08	−5.55	162.37	182.39	167.56	55.56	29.14	37.86	26.41	<b>8.45</b>	0.82	0.75	0.82	<b>0.98</b>
ANN	89.12	100.08	91.97	52.01	4.48	−6.80	8.84	33.36	142.74	154.34	149.64	85.26	25.61	32.04	23.58	12.96	0.86	0.82	0.85	0.95
GA/ANN	<b>62.53</b>	<b>72.89</b>	<b>57.53</b>	<b>37.25</b>	<b>0.23</b>	<b>−0.68</b>	−3.38	7.59	<b>104.28</b>	<b>117.47</b>	<b>98.29</b>	<b>59.12</b>	<b>18.71</b>	<b>24.39</b>	<b>15.49</b>	8.99	<b>0.93</b>	<b>0.89</b>	<b>0.94</b>	<b>0.98</b>

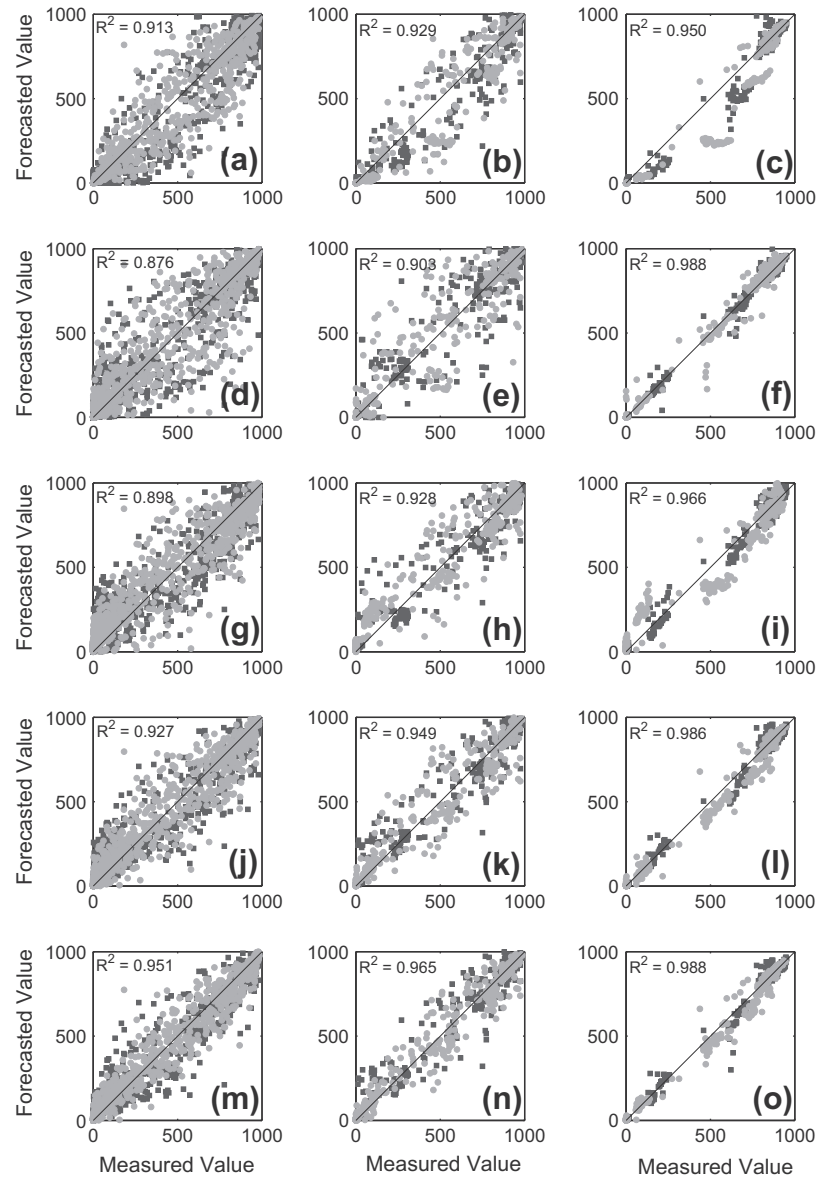


Fig. 8. Scatter plots for the 1 h ahead forecasting. Each row corresponds to a different model. Row 1 (subplots (a)–(c)): persistent model; row 2 (subplots (d)–(f)): kNN model; row 3 (subplots (g)–(i)): ARIMA model, row 4 (subplots (j)–(l)): GA model, row 5 (subplots (m)–(o)): GA/ANN model. Each column corresponds to a different variability period. Left column: forecasting for the period January to April of 2011; Middle: forecasting for the periods May and June of 2011 and Right: July and August of 2011. The symbol ■ identifies morning values and the symbol ● identifies afternoon values.

ferent solutions depending on its initialization. One way to prevent this randomness is to initialize the algorithm to a specific starting point, however given that we have no idea of the topology of the search space for the ANN weights and biases we opt to use the default random initialization. Another way to address this issue is to train the ANN models multiple times and average their output. That is the approach followed here: once the ANN inputs and architecture are defined either manually or through the GA, each ANN is trained 10 times, and the results reported below are the averaged of the 10 different forecasted values.

Fig. 6 exemplifies such operation for the GA optimized ANN for the 1 h ahead forecasting. The forecasted values

obtained with the 10 ANNs are shown together with the averaged forecasted value and the measured value. For a given time stamp the 95% confidence interval is centered on the averaged forecasted value with a width of  $2 \times 2.262$  standard deviations of the 10 forecasted values (where 2.262% is the 95% significance multiplicative factor from the *t-student* distribution with 9° of freedom). The narrow forecasting bands lead us to conclude that despite the randomness in the training of the ANN all the solutions will be very close together and will not deviate much from the average forecasting. As the figure also shows the band will be narrower for sections of the time series that are more well behaved.



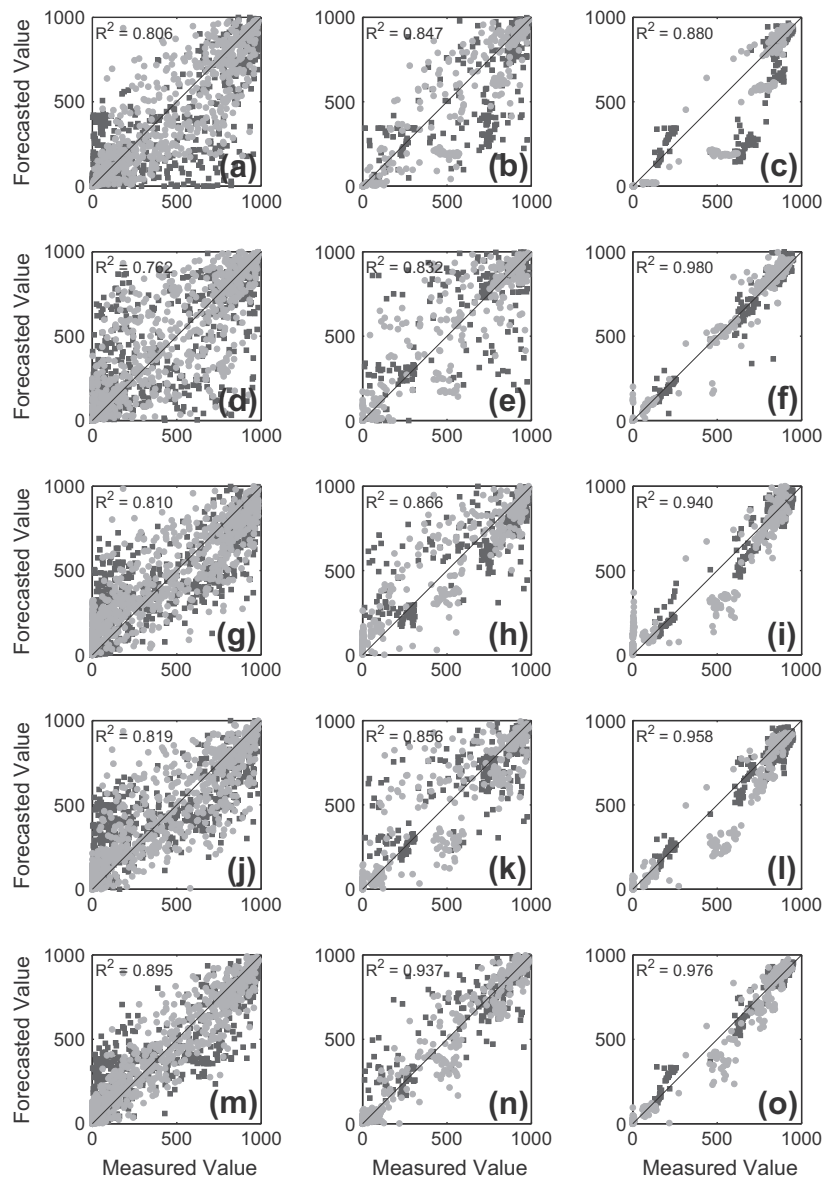


Fig. 9. Scatter plots for the 2 h ahead forecasting. Each row corresponds to a different model. Row 1 (subplots (a)–(c)): persistent model; row 2 (subplots (d)–(f)): kNN model; row 3 (subplots (g)–(i)): ARIMA model, row 4 (subplots (j)–(l)): GA model, row 5 (subplots (m)–(o)): GA/ANN model. Each column corresponds to a different variability period. Left column: forecasting for the period January to April of 2011; Middle: forecasting for the periods May and June of 2011 and Right: July and August of 2011. The symbol ■ identifies morning values and the symbol ● identifies afternoon values.

#### 4.1.1. GA/ANN model

The parameters for the best ANNs models for the 1 h and 2 h forecasting horizons optimized with the GA are given in Table 1. The list of lagged values selected by the GA to include in the ANN input makes sense in the face of the information obtained from the autocorrelation plot: all the higher correlation lags in Fig. 5 are present in Table 1. With respect to the ANN architecture, the selection of three hidden layers with 12 and 16 neurons per layers for the 1 h and 2 h forecasting horizon, respectively, shows that the original architecture of 1 layer and 20 neurons had plenty room for improvement. For more complex models involving multiple variables, such as solar irradiance (GHI, DNI, etc.), weather measurements, cloud

cover, multiple preprocessing techniques, etc., the parameter space for the ANN model will grow very fast. In those cases it is expected that the GA optimization will even more useful in selecting the optimal parameters for the ANN models.

#### 4.2. Comparison of the methods

The models built upon the historical data of 2009 and 2010 (data in the shaded area in Fig. 3) are applied to the 2011 data (data in the unshaded area in Fig. 3) without modifications or retraining. This way it is guaranteed that the data used for training the models is independent of the data used in the evaluation of the models accuracy. After

computing the forecasted power output from the clear sky model and the forecasted stochastic component the models' accuracy is evaluated using the following statistical metrics:

- Mean Absolute Error

$$MAE = \frac{1}{m} \sum_{t=1}^m |P_t - \hat{P}_t| \quad (12)$$

- Mean Bias Error

$$MBE = \frac{1}{m} \sum_{t=1}^m P_t - \hat{P}_t \quad (13)$$

- Root Mean Square Error

$$RMSE = \sqrt{\frac{1}{m} \sum_{t=1}^m (P_t - \hat{P}_t)^2} \quad (14)$$

- normalized Root Mean Square Error

$$nRMSE = \sqrt{\frac{\sum_{t=1}^m (P_t - \hat{P}_t)^2}{\sum_{t=1}^m P_t^2}} \quad (15)$$

- Coefficient of determination

$$R^2 = 1 - \frac{\sum_{t=1}^m (P_t - \hat{P}_t)^2}{\sum_{t=1}^m (P_t - \bar{P})^2} \quad (16)$$

where  $\bar{P} = \frac{1}{m} \sum_{t=1}^m P_t$  and  $m$  is the size of the error evaluation data set,  $P_t$  is the measured 1 h averaged power output and  $\hat{P}_t$  is the 1 or 2 h ahead forecasting for  $P_t$ .

Given that, as seen in Fig. 1, there is a strong seasonality in the variability of the power output, we expect a strong seasonality in the accuracy of the predicted values. In order to study this factor, we consider three seasons, or periods, which are subsets of the total error evaluation data set. The three periods are defined based on the variability study in Fig. 1 as:

- a high variability period from January 1, 2011 to April 30, 2011 identified by “ $P_1$ ”;
- a medium variability period from May 1, 2011 to June 30, 2011 identified by “ $P_2$ ”;
- and a low variability period from July 1, 2011 to August 15, 2011 identified by “ $P_3$ ”.

Fig. 7 shows the probability density function (PDF) for the three variability periods. The three PDFs are very different from each other which validates the data splitting in these three periods for the error analysis.

All the different statistical metrics for the error are also calculated for this three periods. Tables 2 and 3 list all the values for the 1 h and 2 h time horizons, respectively. The headers “ $P_1$ ”, “ $P_2$ ” and “ $P_3$ ” and the header “Tot” identify the error values for the three subsets and for the entire validation data set, respectively. The bold-faced values identify the best model for a given error metric and a given data set.

Besides these quantitative error measures, one of the easiest ways of qualitatively assessing the accuracy of the forecasting is through a scatter plot of the pairs  $(P_t, \hat{P}_t)$ . The better the forecasting the closer the points are from the identity line, and the closer the coefficient of correlation  $R^2$  is to the unity. The scatter plots for the 1 h ahead forecasting are depicted in Fig. 8 and the scatter plots for the 2 h ahead forecasting are shown in Fig. 9. In these figures each row corresponds to a different forecasting model and each column corresponds to a different variability period. In the scatter plots morning values and afternoon values are identified by different symbols. These plots allow us to see the forecasted values are dispersed about the identity line more or less uniformly independently of being a morning forecasting or a afternoon forecasting. This is a clear evidence that the models are free of systematic errors related to the daily solar variation.

The analysis of Tables 2 and 3 shows that the two ANN-based methods, the ANN and the GA/ANN clearly outperform the other models. Only in terms of MBE is the

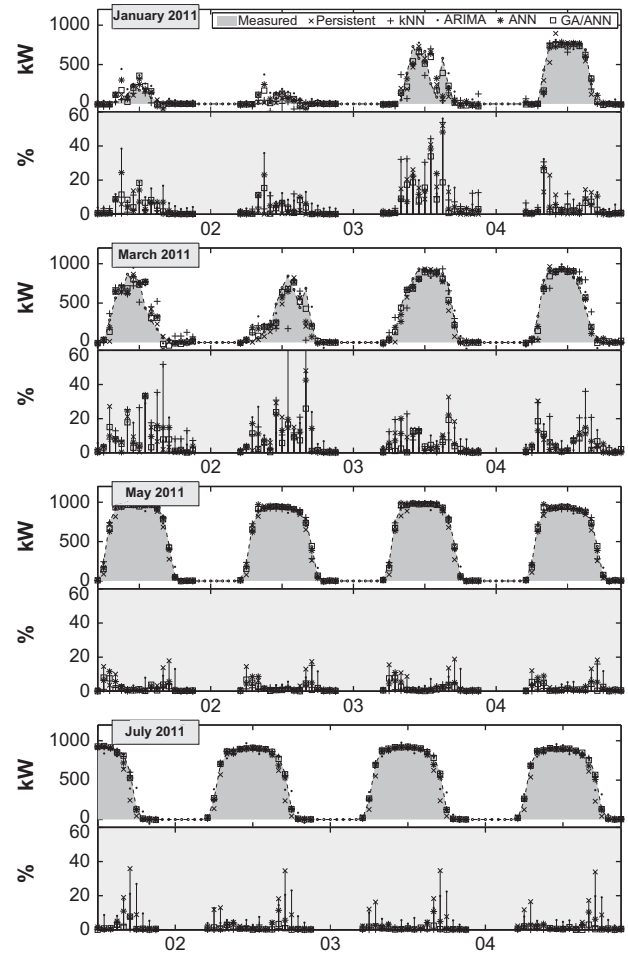


Fig. 10. Comparison between 1 h ahead forecast and measured values of power output for the first 4 days of January, March, May and July of 2011. The point-wise relative error (w.r.t the peak  $P$  of 1000 kW) is shown below each time series plot.

GA/ANN worse than the ARIMA for certain periods. The tables also show how strongly the accuracy of the methods depends on the season. For all models the error metrics for “ $P_3$ ” are substantially better than for the other two periods.

Both tables shown that, for its simplicity the kNN performs very well for low variability situations. This is expected given that in those cases the mapping (pattern, forecasting) becomes “almost” deterministic, on the other hand for the periods of medium and high variability the kNN performs the worst for most error metrics.

The results for the ANN-based models allow us to conclude that there the initial input selection and architecture for the ANN was indeed suboptimal. The results for the GA/ANN represent a big improvement with respect to the results from the ANN predictor for both forecasting horizons. Moreover this improvement is more substantial for the periods of higher variability “ $P_1$ ” and “ $P_2$ ”. The scatterplots (m) and (n) from Fig. 8 show a very clear clustering of the data close to the unity line when compared to the plots (j) and (k). For the 2 h ahead forecasting that improvement is even greater as seen in Fig. 9.

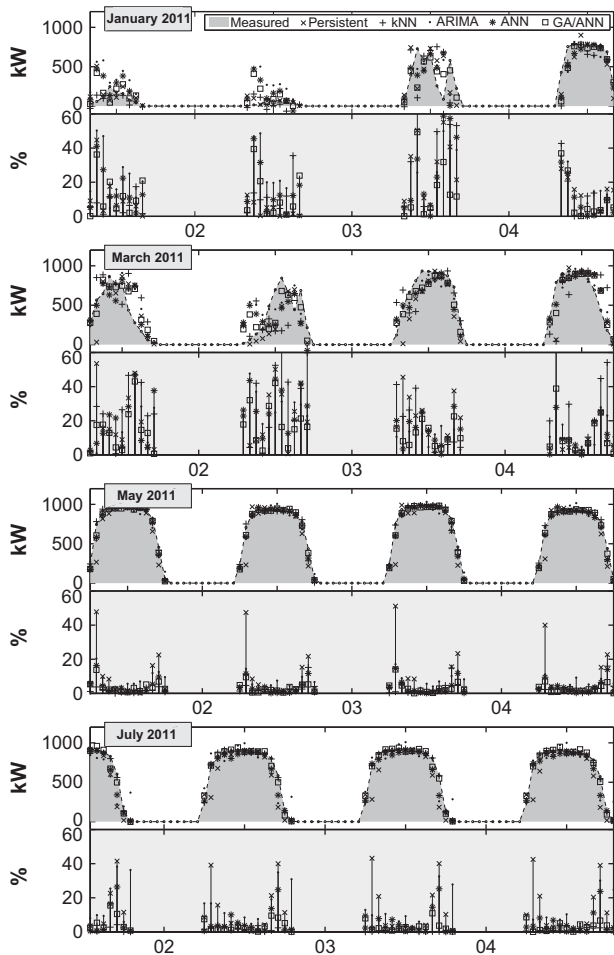


Fig. 11. Comparison between 2 h ahead forecast and measured values of power output for the first 4 days of January, March, May and July of 2011. The point-wise relative error (w.r.t the peak  $P$  of 1000 kW) is shown below each time series plot.

Table 4

Improvement in forecasting skill with respect to the smart persistent model as measured by the decrease of RMSE for the validation data set. Negative values indicate an increase in RMSE.

Persist	Forecasting horizon	
	1 h	2 h
RMSE	107.48 kW	160.79 kW
ARIMA	1.7%	10.3%
KNN	−8.4%	−1.0%
ANN	17.9%	11.2%
GA/ANN	32.2%	35.1%

Figs. 10 and 11 plot the measured and the forecasted time-series for the first 4 days from the months February, April, May and July for the 1 h and the 2 h time-horizon forecasting, respectively. Below each time series, it is plotted the instantaneous absolute relative error (w.r.t. the nominal peak production of 1000 kW).

Finally, Table 4 compares the ARIMA, the kNN and ANN-based models with respect to the persistent model in terms of RMSE for the entire validation period. A positive value indicates a decrease in the RMSE relative to the Persistent model and a negative value indicates a increase in the RMSE. For this table we use the RMSE values for the entire validation set, that is, the RMSE in the column identified by “Tot” in Tables 2 and 3. Table 4 show that overall only the kNN performs worse than the Persistent model. The ARIMA shows a substantial improvement for the 2 h time-horizon. Both ANN-based models outperform every other model with improvements of more than 30% in the case of the GA/ANN model.

## 5. Conclusions

This work assesses five techniques for the 1 h ahead and 2 h ahead forecasting of the averaged power output of a 1 MWp photovoltaic power plant. No exogenous data such as solar irradiance telemetry was used in the forecasting models, which essentially means that the solar panels themselves are the only “sensors” used to generate input data in this work.

From the analysis of the error between measured values and forecasted values it can be concluded that ANN models outperform Persistent, the ARIMA and kNN class models. Moreover, we show that input selection and ANN architecture can be optimized via genetic algorithms in order to improve the accuracy of the ANN forecasting model. For the present case, where the space of inputs is limited due to the lack of exogenous variables, the gains obtained with the GA optimization are already substantial for both time horizons. If other input variables, preprocessing techniques and ANN architectures are included in the list of optimization parameters, one should expect even more expressive gains when using the hybrid GA/ANN methodology.

The analysis of the error as a parameter for determining power output variability implies that the forecasting skill

has a strong seasonal dependency in California's Central Valley, and it suggests that developing seasonal models for each "variability period" (instead of a single model for the year round data) should improve the assessment of the forecasting skill. This is due to the fact that some seasons have too many clear days that are easier to forecast, therefore reducing the differential in errors for all forecasting methods. Nonetheless, our production, variability and forecastability data show that California's Central Valley has a nearly "ideal" seasonal profile for solar production since the lowest variability (and thus highest forecastability) occurs both at the peak load time of the day and at the peak load season.

### Acknowledgements

CFMC and HTCP gratefully acknowledge the partial financial support given by the California Energy Commission (CEC) under the PIER RESCO Project PIR-07-036, and by the National Science Foundation (NSF) CNS division Grant No. 0923586.

### References

- Armano, G., Marchesi, M., Murru, A., 2005. A hybrid genetic-neural architecture for stock indexes forecasting. *Information Sciences* 170, 3–33.
- Bacher, P., Madsen, H., Nielsen, H.A., 2009. Online short-term solar power forecasting. *Solar Energy* 83, 1772–1783.
- Bishop, C.M., 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.
- Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 2008. *Time Series Analysis: Forecasting and Control*, fourth ed. Wiley.
- Brockwell, P.J., Davis, R.A., 2002. *Introduction to Time Series and Forecasting*. Springer, New York.
- Castillo, P., Merelo, J., Prieto, A., Rivas, V., Romero, G., 2000. G-Prop: global optimization of multilayer perceptrons using GAs. *Neurocomputing* 35, 149–163.
- Chen, C., Duan, S., Cai, T., Liu, B., 2011. Online 24-h solar power forecasting based on weather type classification using artificial neural network. *Solar Energy* 85, 2856–2870.
- Chow, C.W., Urquhart, B., Lave, M., Dominguez, A., Kleissl, J., Shields, J., Washom, B., 2011. Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed. *Solar Energy*, 2881–2893.
- Duda, R., Hart, P., 2000. *Pattern Classification*, second ed. John Wiley & Sons.
- Elizondo, D., Hoogenboom, G., McClendon, R., 1994. Development of a neural network model to predict daily solar radiation. *Agricultural and Forest Meteorology* 71, 115–132.
- Gordon, R., 2009. Predicting solar radiation at high resolutions: a comparison of time series forecasts. *Solar Energy* 83, 342–349.
- Hammer, A., Heinemann, D., Lorenz, E., Lücke, B., 1999. Short-term forecasting of solar radiation: a statistical approach using satellite data. *Solar Energy* 67, 139–150.
- Holland, J.H., 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor MI.
- Lara-Fanego, V., Ruiz-Arias, J., Pozo-Vázquez, D., Santos-Alamillos, F., Tovar-Pescador, J., 2011. Evaluation of the WRF model solar irradiance forecasts in andalusia (Southern Spain). *Solar Energy*.
- Marquez, R., Coimbra, C.F.M., 2011. Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database. *Solar Energy* 85, 746–756.
- Martin, L., Zarzalejo, L.F., Polo, J., Navarro, A., Marchante, R., Cony, M., 2010. Prediction of global solar irradiance based on time series analysis: application to solar thermal power plants energy production planning. *Solar Energy* 84, 1772–1781.
- Mellit, A., 2008. Artificial intelligence technique for modelling and forecasting of solar radiation data: a review. *International Journal of Artificial Intelligence and Soft Computing* 1, 52–76.
- Mellit, A., Pavan, A.M., 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. *Solar Energy* 84, 807–821.
- Mohandes, M., Rehman, S., Halawani, T.O., 1998. Estimation of global solar radiation using artificial neural networks. *Renewable Energy* 14, 179–184.
- Paoli, C., Voyant, C., Muselli, M., Nivet, M., 2010. Forecasting of preprocessed daily solar radiation time series using neural networks. *Solar Energy* 84, 2146–2160.
- Picault, D., Raison, B., Bacha, S., de la Casa, J., Aguilera, J., 2010. Forecasting photovoltaic array power production subject to mismatch losses. *Solar Energy* 84, 1301–1309.
- Sfetsos, A., Coonick, A., 2000. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. *Solar Energy* 68, 169–178.
- Zhang, G.Q., Patuwo, B.E., Hu, M.Y., 1998. Forecasting with artificial neural networks: the state of the art. *International Journal of Forecasting* 14, 35–62.