USING A NONPARAMETRIC PV MODEL TO FORECAST AC POWER OUTPUT OF PV PLANTS

Marcelo Pinho Almeida[a], Oscar Perpiñán[b,c], Luis Narvarte[c]

[a] Instituto de Energia e Ambiente
Universidade de São Paulo
Avenida Professor Luciano Gualberto 1289, Cidade Universitária, 05508-010 São Paulo, Brazil

[b] Departamento de Ingeniería Eléctrica, Electrónica, Automática y Física Aplicada
Escuela Técnica Superior de Ingeniería y Diseño Industrial
Universidad Politécnica de Madrid
Ronda de Valencia 3, 28012 Madrid, Spain

[c] Instituto de Energía Solar
Universidad Politécnica de Madrid
Ciudad Universitaria s/n, Madrid, Spain

E-mail: marcelopa@iee.usp.br

ABSTRACT: In this paper, a methodology using a nonparametric model is used to forecast AC power output of PV plants using as inputs several forecasts of meteorological variables from a Numerical Weather Prediction (NWP) model and actual AC power measurements of PV plants. The methodology was built upon the R environment and uses Quantile Regression Forests as machine learning tool to forecast the AC power with a confidence interval. Real data from five PV plants was used to validate the methodology, and results show that the daily production of individual plants can be predicted with a skill score up to 0.361.
Keywords: PV output power forecast, Numerical Weather Prediction, Quantile Regression Forests

1 INTRODUCTION

An accurate AC power output forecast of PV plants is an important matter for both plant owners and electric system operators. This paper conceives the PV system as a black box, presuming no knowledge of any internal characteristics and processes of the system, treating it as a data-driven model that estimates the behavior of the system from a historical time series of inputs and outputs.

This nonparametric approach circumvents the need for simplifying assumptions and accurate internal parameters with the use of historical time series of meteorological variables and AC power measurements. Therefore, its accuracy depends mainly on the quality of the input data. One interesting advantage of a nonparametric model is the potential to compensate systematic errors associated to the inputs.

The nonparametric approach has been implemented in several recent researches. Bacher et al. [1] forecasts hourly values of AC power of PV systems for horizons of up to 36 h using adaptive linear time series models using Numerical Weather Predictions as input. Mandal et al. [2] forecasts one-hour-ahead power output of a PV system using a combination of wavelet transform and neural network techniques by incorporating the interactions of PV system with solar radiation and temperature data. Pedro and Coimbra [3] predicts 1 and 2 h-ahead solar power of a PV system comparing several forecasting techniques without exogenous inputs such as Auto-Regressive Integrated Moving Average, k-Nearest-Neighbors, Artificial Neural Networks, and Neural Networks optimized by Genetic Algorithms. Zamo et al. [4] analyzes a mix of eight statistical methods to forecast PV power one day ahead in an hourly basis, and the Random Forests method presents the best results.

This paper proposes a methodology to derive AC power forecasts one day ahead with hourly resolution using a nonparametric PV model based on Quantile Regression Forests. Both a single-valued forecast and a probabilistic forecast are produced, providing statistical information about the uncertainty of the output. Several variability indexes derived from the original variables are proposed, and a systematic and exhaustive variable importance analysis is carried out with different scenarios. The length of the time series used to learn from data, as well as the method for selecting the days included in this training time series, are analyzed regarding the model's performance.

The methodology has been validated by comparing the predictions with measured AC power from several PV plants, as described in Section 4. The results are presented in Sections 5 and 6.

2 DESCRIPTION OF THE METHODOLOGY

The proposed methodology is as follows:

2.1 Gathering data

Previous AC power measurements from a PV plant and forecasts of a set of Weather Research and Forecasting (WRF) variables (solar radiation, cloud cover, temperature, wind speed, etc.) from a Numerical Weather Prediction (NWP) model are collected.

The database of real AC power used in this paper comes from five PV plants situated in northern Spain (latitude ≈42.2°), with a 5-s resolution, previously analyzed in [5]. The data is comprised between January 1st, 2009 and December 29th, 2010. Moreover, in order to reduce file sizes and to filter noise, the raw data has been aggregated to produce 1-min records, which was then aggregated into 1-h values.

Table I summarizes the main characteristics of these PV plants. Their installed power ranges from 2.64 MWp to 958 kWp, with areas ranging from 11.8 ha to 4.1 ha. They all have an azimuthal one-axis tracker, with a receiving surface tilted 45°.

Table I: PV plants characteristics

Label	Peak power (kWp)	Rated Power (kW)	Area (Ha)
P1	958	775	4.1
P2	990	780	4.2
P3	1438	1155	6.4
P4	1780	1400	8.7
P5	2640	2000	11.8

The WRF variables are downloaded from Meteogalicia, a meteorological institute of the Xunta de Galicia (Spain), which regularly publishes results from a regional mesoscale NWP model, the Weather Research and Forecasting [6], freely at its THREDDS server. The model runs twice a day, initialized at 00UTC (forecast for the next 96 hours) and 12UTC (forecast for the next 84 hours). The spatial resolution is 12 km x 12 km, in an area comprised between 21.58W to 6.36E and 33.64N to 49.57N, and the temporal resolution is hourly. Meteogalicia also maintains an historical archive of past forecasts that are available online. Table II presents the name and the description of the WRF variables considered in this paper.

Table II: WRF-NWP variables used to forecast AC power in this paper

Label	Description
swflx	surface down welling shortwave flux
temp	temperature at 2 m
cfh	cloud cover at high levels
cfl	cloud cover at low levels
cfm	cloud cover at mid-levels
cft	cloud cover at low and mid-levels
u	longitude-wind at 10 m
v	latitude-wind at 10 m
mod	wind speed module at 10 m
dir	wind direction at 10 m
rh	relative humidity at 2 m
mslp	mean sea level pressure
visibility	visibility in air

2.2 Processing the WRF variables

Each WRF variable is processed to extract information about the value at the location of interest and its relation with the surrounding locations and several previous forecasts.

Both the actual values of the WRF variables (*point*) and the interpolated values of adjacent locations using Inverse Distance Interpolation (*IDW*) are computed for the location of interest. Fig. 1 displays forecasts of solar irradiance for several nearby locations around PV plant P1. Fig. 2 shows the forecast of solar irradiance produced by several consecutive model runs for the position of PV plant P1. During clear sky or complete cloudy days, forecasts of different model runs and at different nearby locations are similar. However, during partially cloudy days, forecasts vary both spatially (different locations) and temporally (different model runs).

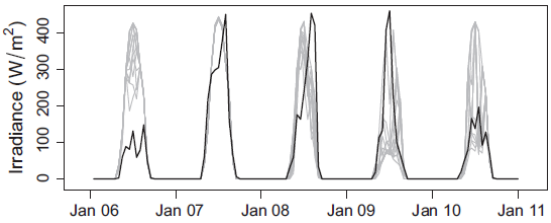

Figure 1: Global horizontal irradiance forecasts for several nearby locations around PV plant P1, and comparison with on-ground measurements (dark line)

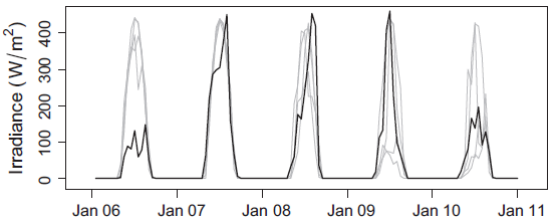

Figure 2: Global horizontal irradiance forecasts produced by several consecutive model runs for the location of PV plant P1, and comparison with on-ground measurements (dark line)

To give information about this variability of the meteorological forecasts, four derived spatial and time indexes are added, as described in Table III.

Table III: Spatial and time variability indexes

Index	Description
TRI	Terrain Ruggedness Index, is defined as the mean of the absolute differences between a central cell and its surrounding 8 cells in a 3 by 3 grid
TPI	Topographic Position Index, is defined as the difference between a central cell and the mean of its surrounding 8 cells in a 3 by 3 grid
roughness	is the largest inter-cell difference of a central cell and its surrounding 8 cells in a 3 by 3 grid
sdr	is the standard deviation of a collection of consecutive model runs

2.3 Preparing train and test time series

In addition to the actual (*point*) and interpolated (*IDW*) values of the WRF variables and the variability indexes (*TRI*, *TPI*, *roughness* and *sdr*), variables that describe the Sun-Earth geometry, azimuth angle (*AzS*), altitude angle (*AlS*) and extra-terrestrial horizontal irradiance (*Bo0*), and the hourly clearness index (*kt*) are also calculated. Along with previous AC power measurements, they compose the set of predictors used in the proposed methodology.

The predictors are divided into two independent time series: train and test. The train time series comprises past values of processed WRF variables and AC power, whereas the test time series contains only present processed WRF variables (forecasts).

The train time series, or training set, may have different sizes, what eventually leads to distinct results. For practical purposes, the size of the training set is defined here in days. Therefore, the training set can be composed by *N* days, selected from a bigger database.

Table IV presents the three selecting methods that were analyzed.

Table IV: Selecting methods of *N* days from the database

Index	Description
Previous	This method selects those *N* days immediately before the day to be predicted. As a consequence, the database must be complete up to the day prior the prediction
KT	This method selects *N* days according to the absolute difference between the clearness index of the day to be predicted and the clearness index of each day included in the database. Both clearness indexes are computed with the irradiance forecast retrieved from the NWP model. The *N* days with the lowest absolute difference are chosen to conform the training set
KS	This method selects *N* days according to the similarity between the empirical distribution function of the irradiance forecast for the day to be predicted and the empirical distribution function of the irradiance forecast for each day included in the database. Here the Kolmogorov-Smirnov statistic is used to compute the distance between the distributions. The *N* days with the lowest Kolmogorov-Smirnov distance are chosen to conform the training set

Both selecting methods *KT* and *KS* do not need the database to be completed up to the day prior the prediction, and could also be composed by older information.

2.4 The machine learning tool

Random Forests [7] are a machine learning tool which consists of a collection, or ensemble, of a multitude of decision trees, each one built from a sample drawn with replacement (a bootstrap sample) from a training set (X_i, Y_i), i = 1, 2,…, where X is the group of inputs and Y is the group of outputs. In addition, when splitting a node during the construction of a tree, only a random subset of variables is used. As a consequence, the final nodes, or leafs, may contain none or several observations from Y.

While Random Forests keeps only an average of the observations of Y that fall into each leaf of each tree and neglects all other information, Quantile Regression Forests keeps the values of all observations of Y in every leaf and assesses the conditional distribution based on this information, enabling the construction of prediction intervals [8].

2.5 Predicting AC output power

The machine learning tool is trained with the train time series. Predictions of the median (quantile 0.5) and the confidence interval (quantiles 0.1 and 0.9) for the AC output power are generated with the test time series. Fig. 3 shows an example of simulation using the proposed methodology with N = 30 days and selecting method *KS*.

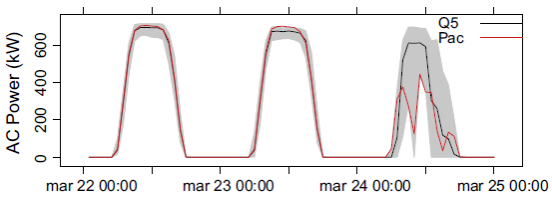

Figure 3: Example of simulation with N = 30 days and selecting method *KS*. The grey area corresponds to the confidence interval

3 TOOLBOX

An online toolbox to implement the methodology was built upon the R environment using a list of contributed packages:

- *rgdal* and *raster* for raster data manipulation [9, 10].
- *zoo*, *xts*, and *data.table* for time series analysis [11, 12, 13].
- *gstat* for spatial interpolation [14].
- *meteoForecast* to import NWP-WRF forecasts [15].
- *solaR* for sun geometry calculation [16].
- *quantregforest* for Quantile Regression Forests [8].
- *PVF* for AC power prediction. This package resumes the implementation of the methodology [17].

Packages *meteoForecast* and *PVF* were developed during the present study. The toolbox is freely available at http://vps156.cesvima.upm.es:3838/predictPac/.

4 VALIDATION

A model performance is commonly evaluated by quantifying the discrepancy between forecasts and actual observations through the use of different statistics such as the Mean Bias Error (*MBE*), the Root Mean Square Error (*RMSE*) and the Mean Absolute Error (*MAE*) [18]. Because each performance statistic characterizes a certain aspect of the overall model performance, a complete evaluation needs the combined use of a collection of these statistics tools.

On the other hand, the proposed methodology produces both the forecast of the median ($Q_{.5}$) and the forecast of the confidence interval between the quantiles $Q_{.1}$ and $Q_{.9}$. To assess the amplitude of the confidence interval, its area, normalized respect to the area of the observations, $Q1Q9_{Sum}$, is calculated with Eq. 1.

$$Q1Q9_{Sum} = \frac{\sum_{i=1}^{n}(Q_{.9i} - Q_{.1i})}{\sum_{i=1}^{n} o_i} \quad (1)$$

where n is equal to 24 hours, as each statistic is computed for a day and the predictions are made on an hourly basis. This statistic gives information on how wide the interval is, as well as how many times the area (or energy) inside the interval is bigger than that comprised under the observed power curve, so greater values of $Q1Q9_{Sum}$ means more uncertainty related to the quantile $Q_{.5}$.

The performance statistic for the accuracy of quantile $Q_{.1}$ considers only the moments (or hours, specifically for this study) when the observed value is smaller than the quantile. First, these moments are identified using Eq. 2, resulting in the vector $Q1u$ (the numbers 0 and 1 were randomly chosen, but simplify the next step). The sum of all elements of $Q1u$ results in the number of exceedances of the observed value regarding $Q_{.1}$ in the period

considered, resulting in $Q1_{Num}$, given by Eq. 3.

$$Q1u_i = \begin{cases} 1 & Q_{.1i} - o_i > 0 \\ 0 & Q_{.1i} - o_i \leqslant 0 \end{cases} \text{ where } i = 1, 2, \ldots, n \quad (2)$$

$$Q1_{Num} = \sum_{i=1}^{n} Q1u_i \quad (3)$$

where n is, again, equal to 24. A similar approach is used to compute the statistic for the accuracy of quantile $Q_{.9}$, but now considering the moments when the observations are greater than the quantile, as described in Eq. 4 and Eq. 5.

$$Q9u_i = \begin{cases} 1 & o_i - Q_{.9i} > 0 \\ 0 & o_i - Q_{.9i} \leqslant 0 \end{cases} \text{ where } i = 1, 2, \ldots, n \quad (4)$$

$$Q9_{Num} = \sum_{i=1}^{n} Q9u_i \quad (5)$$

The performance of the proposed methodology was using a leave-one-out cross-validation procedure:

- One day is extracted from the database to be the test set. The AC power measurements (observations) are stored separately.
- The training set is constructed with N days extracted from the remaining days of the data set, according to the selecting method (*KS*; *KT*, or *Previous*). This training set is used to train the machine learning tool.
- Predictions AC power, with hourly quantiles $Q_{.1}$; $Q_{.5}$ and $Q_{.9}$, are obtained for the test set.
- Using the quantiles from the step above and the previously stored observations, the performance statistics are calculated. Each day from the database is then characterized by several performance statistics: *RMSE*, *MBE*, *MAE*, *Q1Q9Sum*, $Q1_{Num}$ and $Q9_{Num}$.

This procedure was repeated for every day in the dataset (over 600 days for each PV plant) and the simulations were made for 17 scenarios, the 3 training set selecting methods and 5 different N (7, 15, 30, 45 and 60 days), resulting in a massive collection of performance statistics.

For ease of understanding, the results of each performance statistic have been aggregated with the quantiles 0.25, 0.5 and 0.75, hereafter denominated $QS_{.25}$; $QS_{.5}$ and $QS_{.75}$, respectively, to distinguish them from the quantiles of the predictions.

Moreover, the model's performance was compared with a persistence method commonly used as reference in forecast problems related with PV generation. This comparison was evaluated using the skill score, defined in Eq. 6.

$$SS = 1 - \frac{RMSE_f}{RMSE_p} \quad (6)$$

where the index f stands for the proposed methodology and p stands for the persistence method.

To make comparison between simulations easier, *MBE*; *RMSE* and *MAE* have been normalized in order to fall in a more restricted range of values. In statistic studies, it is common to normalize these statistics to the range, *max(**O**)* – *min(**O**)*, or the mean, *mean(**O**)*, of the observations (**O**). Table V summarizes the performance statistics used. The first option was chosen for a statistical analysis to ensure most of the values fall in a range between 0 and 1.

Table V: Performance statistics

	Index	Description
Quantile $Q_{.5}$	*nMBE*	*MBE* normalized respect to the daily range of the observations
	nRMSE	*RMSE* normalized respect to the daily range of the observations
	nMAE	*MAE* normalized respect to the daily range of the observations
	SS	As it is (calculated with *nRMSE*)
Confidence interval	$Q1Q9_{Sum}$	As it is
	$Q1_{Num}$	As it is
	$Q9_{Num}$	As it is

The results are grouped accordingly to the daily clearness index (*KTd*) into three classes: cloudy days ($0 \leq KTd < 0.532$), partially cloudy days ($0.532 \leq KTd < 0.678$) and clear days ($0.678 \leq KTd < 1$). The ranges of *KTd* were selected so that the classes comprise one third of the total number of days present in the database.

The full picture of the strengths and weakness of a complex model is only grasped when its performance is evaluated under different conditions. The proposed methodology was examined varying the predictors used. A total of 17 scenarios were defined to analyze the model performance. Table VI summarizes the characteristics of these scenarios.

5 STATISTICAL RESULTS

For practical purposes only the detailed results for P1 are presented here, as they represent the overall behavior of the other four PV plants. From a preliminary analysis of these results, it can be stated that the performance is almost independent of the scenario chosen, as long as it contains irradiance as input, either *swflx* WRF variable or calculated extraterrestrial irradiance, *Bo0*. Fig. 4, for N = 30 days, illustrate this behavior:

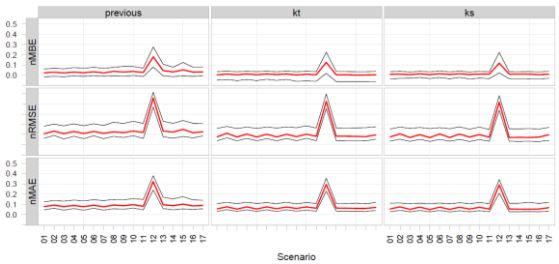

Figure 4: Statistics for the quantile $Q_{.5}$ for N = 30 days and $0.678 \leq KTd \leq 1$

It is clear that scenario 12, which does not incorporates any direct information about irradiance, presents the worst performance. Therefore, it is possible to state that irradiance data must be present in the predictors and a large collection of WRF variables is not mandatory, so if only a few are available, the performance of the forecasts will not be compromised. A similar result was obtained in [4].

Table VI: A total of 17 scenarios were defined to analyze the non-parametric model performance. They differ on the variables and indexes used as predictors

Scenario		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
WRF variables (*point* and *IDW*)	swflx	×	×	×	×	×	×	×	×	×	×						×	×
	temp	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	cft	×	×	×	×	×	×					×	×		×			
	cfl	×	×	×	×							×	×					
	cfm	×	×	×	×							×	×					
	cfh	×	×	×	×							×	×					
	u	×	×	×	×	×	×											
	v	×	×	×	×	×	×											
	mod	×	×	×	×	×	×	×	×	×	×	×	×	×	×		×	×
	dir	×	×	×	×	×	×	×	×									
	rh	×	×	×	×	×	×	×	×	×	×	×	×	×	×		×	×
	visibility	×	×	×	×	×	×	×	×									
	mslp	×	×	×	×	×	×	×	×									
Sun-Earth Geometry	*AlS*	×		×		×		×		×		×		×	×	×	×	
	AzS	×		×		×		×		×		×		×	×	×	×	
	Bo0	×		×		×		×		×		×		×	×	×	×	
Indexes (for WRF variables)	*TRI*	×	×															
	TPI	×	×															
	rough	×	×															
	sdr	×	×															
Clearness Index	*kt*																×	×

From this point on, only the results for scenarios 1, the most complete regarding WRF variables, indexes and computed variables, and 9, which has the smallest number of variables in the predictor set under the condition of containing both predicted (*swflx*) and calculated (*Bo0*) irradiances, will be detailed.

Fig. 5 ranks the scenarios according to their Skill Score performances.

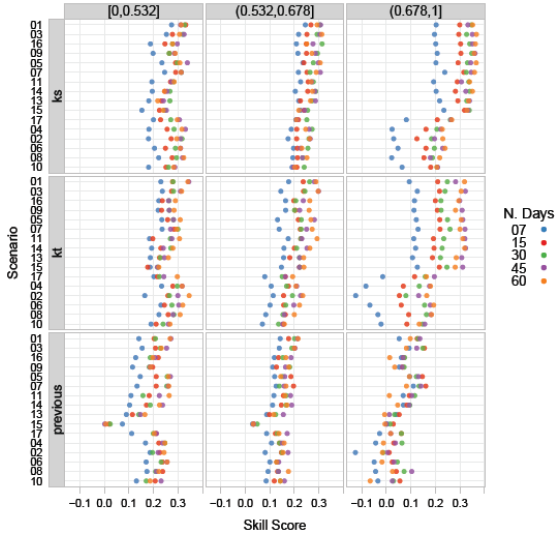

Figure 5: Median of the daily skill scores for each scenario considering the possible combinations of *KTd*, *N*, and selecting method. Scenarios are ordered according to their skill score performance. Results for scenario 12 are not presented due to its lower performance (*SS* between -0.07 and -2.04)

Fig. 5 shows that the best selecting method is *KS* and the influence of *N* is inappreciable when $N > 30$ days. Therefore, only the results with selecting method *KS* and $N = 30$ days will be presented from now on.

The values of *SS* for scenarios 1 and 9 are 0.346 and 0.348, respectively. These results compare satisfactorily with those reported in [1] with a set of forecast methods of AC power for next day horizons. These authors published skill scores up to 0.36 for aggregated forecasts corresponding to the average power of a set of 21 different PV systems in a region. In contrast, our proposal is focused on the forecast of individual PV plants. Only to further illustrate the *SS* performance of the proposed methodology, considering all PV plants from Table I the range of *SS* for scenarios 1 and 9 are 0.336 to 0.361 and 0.324 to 0.350, respectively.

The quantiles $QS_{.25}$, $QS_{.5}$ and $QS_{.75}$ of the performance statistics are presented in Tables VII and VIII for scenarios 1 and 9, respectively. Results are again very similar, reinforcing that as long as irradiance data is present in the predictors, a large collection of WRF variables is not mandatory.

The statistics are computed for a period of one day, so 24 individual hourly errors are resumed into one single value. The *nMBE* indicates a mean accumulated error for the entire period, while the *nRMSE* and *nMAE* give some insight on the individual errors.

The median ($QS_{.5}$) *nMBE* is small for both scenarios and all *KTd* classes, with a maximum of 4%. This is expected from a statistical method based on Random Forests, which has the tendency to give unbiased results. Individual errors are somewhat bigger, as can be observed from the higher values of *nRMSE* and *nMAE*. Nevertheless, for clear days, which concentrate most of the electricity generation (almost 50%), these statistics are still very good.

The uncertainty related to the quantile $Q_{.5}$ is relatively low for clear and partially clouded days. For cloudy days, $Q1Q9_{Sum}$ indicates a higher level of uncertainty, but this is strictly related to the variability of the solar resource due to unstable cloud cover and the small amount of energy generated during cloudy days, which is closer to the magnitude of the generation uncertainty.

Statistical methods based on Random Forests, due to their inherent averaging, tend to avoid minimums and maximums. Therefore, extrapolations regarding the

quantile $Q_{.9}$ are more likely to happen. As the confidence interval is limited by the quantiles $Q_{.1}$ and $Q_{.9}$ and it is calculated for one day (24 hours), no more than 2.4 extrapolations are expected under or over the confidence interval. Median $Q1_{Num}$ and $Q9_{Num}$ are consistent with what was expected. Only with cloudy days and scenario 9 $Q9_{Num}$ presented a number of extrapolations bigger than 2.4 in the range between quantiles $QS_{.25}$ and $QS_{.75}$.

Table VII: Quantiles $QS_{.25}$, $QS_{.5}$ and $QS_{.75}$ of the performance statistics for each *KTd* class, with N = 30 days, selecting method *KS* and scenario 1

Statistic	$0 \leqslant KTd < 0.532$			$0.532 \leqslant KTd < 0.678$			$0.678 \leqslant KTd \leqslant 1$		
	$QS_{.25}$	$QS_{.5}$	$QS_{.75}$	$QS_{.25}$	$QS_{.5}$	$QS_{.75}$	$QS_{.25}$	$QS_{.5}$	$QS_{.75}$
$nMBE$	9.88%	−4.02%	−18.49%	8.11%	−0.92%	−7.26%	3.74%	−0.75%	−3.28%
$nRMSE$	26.21%	31.65%	42.95%	13.74%	20.90%	27.68%	3.71%	7.71%	15.49%
$nMAE$	19.38%	24.19%	34.64%	10.61%	16.25%	22.05%	2.80%	5.48%	11.13%
$Q1Q9_{Sum}$	1.35	1.99	3.60	0.73	0.99	1.32	0.27	0.36	0.51
$Q1_{Num}$	0	0	1	0	0	2	0	0	1
$Q9_{Num}$	0	1	2	0	1	2	0	1	2

Table VIII: Quantiles $QS_{.25}$, $QS_{.5}$ and $QS_{.75}$ of the performance statistics for each *KTd* class, with N = 30 days, selecting method *KS* and scenario 9

Statistic	$0 \leqslant KTd < 0.532$			$0.532 \leqslant KTd < 0.678$			$0.678 \leqslant KTd \leqslant 1$		
	$QS_{.25}$	$QS_{.5}$	$QS_{.75}$	$QS_{.25}$	$QS_{.5}$	$QS_{.75}$	$QS_{.25}$	$QS_{.5}$	$QS_{.75}$
$nMBE$	17.83%	−1.65%	−19.46%	9.70%	1.98%	−6.35%	3.70%	−0.66%	−3.03%
$nRMSE$	27.03%	35.61%	47.82%	14.13%	21.03%	29.27%	3.29%	7.62%	15.24%
$nMAE$	19.80%	26.53%	39.07%	10.52%	16.13%	22.69%	2.66%	5.36%	10.80%
$Q1Q9_{Sum}$	1.24	1.98	3.62	0.70	0.95	1.22	0.22	0.34	0.49
$Q1_{Num}$	0	0	2	0	0	2	0	0	2
$Q9_{Num}$	0	1	3	0	1	2	0	1	2

6 IMPACTS ON DAILY ENERGY FORECAST

Previous sections have evaluated the methodology's performance under a statistical framework using tools and metrics commonly found in this discipline. However, PV power forecasting can be used for trading energy in electricity power markets.

This section considers this framework, taking into account the economic benefits and penalties stated in the market regulations. There is a variety of market practices and regulations that provokes that a certain forecasting model can perform better or worse due to the different impact of the success and failures in each market. Therefore, the metrics used to evaluate the model performance, in terms of the quantile $Q_{.5}$ of the forecast, must be adequate to the market configuration.

Two important scenarios are accounted here: on the one hand, markets that penalize the daily energy error for which the *MBE* is appropriate; on the other hand, markets that penalize the hourly energy error, for which the *MAE* is preferred. In this context, these metrics are more useful if presented as an energy ratio, and thus they were normalized to the daily measured energy, resulting in *cvMBE* and *cvMAE*, respectively.

The *cvMAE* measures the goodness of the predictions for applications requiring hourly predictions during a period of a day, whereas the *cvMBE* is an index of the goodness of the total daily energy production. Both have been computed for each day included in the database and the median of the results has been calculated. Besides, this median was weighted with the energy generated by the PV plant under the corresponding *KTd* class.

The statistics have been computed with N = 30 days, selecting method *KS* and scenarios 1 and 9, as presented in Table IX.

Table IX: Weighted errors of energy forecast for PV plant P1 according to the *KTd* class, with N = 30 days, selecting method *KS* and scenarios 1 and 9.

KTd class	*cvMBE*		*cvMAE*	
	Sc. 1	Sc. 9	Sc. 1	Sc. 9
$0.00 \leq KTd < 0.53$	-1.27	-0.51	8.63	9.49
$0.53 \leq KTd < 0.68$	-0.47	1.22	9.14	8.76
$0.68 \leq KTd \leq 1.00$	-0.54	-0.49	4.13	4.22

Values of *cvMBE* are small, what was expected due to the machine learning tool used. Total daily energy is forecasted with an absolute *cvMBE* of less than 1.3% for all *KTd* classes. In terms of hourly prediction, the overall *cvMAE* is less than 9.5%. Both results are very good and appear to be independent of the size of the PV plant.

8 CONCLUSION

A methodology to forecast one day ahead hourly AC power produced by a PV plant has been proposed. This approach conceives the PV system as a black box (nonparametric PV model), and it does not presume any knowledge of internal characteristics and processes of the system.

The methodology uses forecasts of several meteorological variables (produced by a Numerical Weather Prediction model), and spatial and temporal indexes (estimated from the forecasted variables) as inputs to predict the hourly AC power of the PV plant. The PV model uses Quantile Regression Forests, which is

able to produce both a central forecast (median) and a confidence interval.

The validation procedure has analyzed the performance of the methodology according to the daily clearness index, the training set length (*N*), the WRF variables and indexes used, and the training set selecting method. The main observations are:

- The presence of irradiance data, predicted (*swflx*) and/or calculated (*Bo0*), leads to better results.
- Increasing the number of WRF variables do not necessarily increase the accuracy of the forecast.
- Training set selecting methods based on similarity measures (*KT* and *KS*) behave better than choosing recent days (*Previous*). Method *KS* achieves the best results.
- The training set length has no significant impact on the model performance with time series longer than 15 days. A value of $N = 30$ days was used with good results.
- The confidence interval ability to contain all observations within is very good, especially for the quantile $Q_{.1}$.
- Total daily energy is forecast with an absolute *cvMBE* of less than 1.3% for all *KTd* classes.
- In terms of hourly prediction, the overall *cvMAE* is less than 9.5%.

The methodology's performance has also been evaluated using the skill score as a measure of the relative improvement over the persistence forecast. The results range from 0.33 to 0.36, comparing satisfactorily with the set of forecast methods reported in [1]. These authors published skill scores up to 0.36 for aggregated forecasts corresponding to the average power of a set of 21 different PV systems, and our proposal is focused on the forecast of individual PV plants.

This paper describes the overall results of an extensive study that is fully detailed in [19], where a rigorous description and validation of the proposed methodology can be found.

ACKNOWLEDGEMENTS

This work has been partially financed by the Seventh Framework Programme of the European Commission with the Project Photovoltaic Cost Reduction, Reliability, Operational Performance, Prediction and Simulation (PVCROPS—Grant Agreement No. 308468).

REFERENCES

[1] P. Bacher, H. Madsen, H. A. Nielsen, Online short-term solar power forecasting. Solar Energy, 83 (2009) 1772–1783.

[2] P. Mandal, S. T. S. Madhirab, A. Ul haquec, J. Mengc, R. L. Pinedaa, Forecasting Power Output of Solar Photovoltaic System Using Wavelet Transform and Artificial Intelligence Techniques. Procedia Computer Science, 12 (2012) 332–337.

[3] H. T. C. Pedro, C. F. M. Coimbra, Assessment of forecasting techniques for solar power production with no exogenous inputs. Solar Energy, 86 (2012) 2017–2028.

[4] M. Zamoa, O. Mestrea, P. Arbogastb, O. Pannekoucke, A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production. Solar Energy, 105 (2014) 792–803.

[5] J. Marcos, L. Marroyo, E. Lorenzo, D. Alvira, E. Izco, Power output fluctuations in large scale PV plants: one year observations with 1 second resolution and a derived analytic model. Prog. In Photovolt.:Res. Appl., 19 (2011) 218–227.

[6] J. B. Klemp, A Description of the Advanced Research WRF Version 2, (2005). Available at: http://www2.mmm.ucar.edu/wrf/users/docs/arw_v2.pdf.

[7] L. E. O. Breiman, Random Forests. (2001) 5–32.

[8] N. Meinshausen, Quantile Regression Forests. The journal of Machine Learning, 7 (2006) 983–999.

[9] R. Bivand, T. Keitt, B. Rowlingson, Bindings for the Geospatial Data Abstraction Library. R package version 0.8-11. (2013). Available at: http://cran.r-project.org/package=rgdal.

[10] R. J. Hijmans, Raster: Geographic Analysis and Modeling with Raster Data. R package version 2.1-66 (2013). Available at: http://cran.r-project.org/package=raster.

[11] M. Dowle, T. Short, S. Lianoglou, A. Srinivasan, Data.table: Extension of data.frame. R package version 1.9.2 (2014). Available at: http:// cran.r - project.org/package=data.table.

[12] J. A. Ryan, J. M. Ulrich, Xts: eXtensible Time Series. R package version 0.9-5e (2013). Available at: http://cran.r-project.org/package=xts.

[13] A. Zeiles, G. Grothendieck, S3 Infraestructure for Regular and Irregular Time Series. Journal of Statistical Software, 14 (2005) 27.

[14] E. J. Pebesma, Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30 (2004) 683–691.

[15] O. Perpiñán, M. P. Almeida, MeteoForecast. R package version 0.31 (2014). Available at: http://dx.doi.org/10.5281/zenodo.10781.

[16] O. Perpiñán, solaR : Solar Radiation and Photovoltaic Systems. Journal of Statistical Software, 50 (2012) 32.

[17] M. P. Almeida, O. Perpiñán, PVF. R package version 0.20 (2014). Available at: http://dx.doi.org/10.5281/zenodo.13348.

[18] C. Gueymard, A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects. Renewable and Sustainable Energy Reviews, 39 (2014) 1024–1034.

[19] M. P. Almeida, O. Perpiñán, L. Narvarte, PV power forecast using a nonparametric PV model. Solar Energy 115 (2015) 354 – 368.