Use of Multilinear Adaptive Regression Splines and Numerical Weather Prediction to forecast the power output of a PV plant in Borkum, Germany

Luca Massidda[a,*], Marino Marrocu[a]

[a] *CRS4, Center for advanced studies, research and development in Sardinia*

Abstract

The development of accurate forecasting methods for renewable energy sources can be an important tool for the integration of such systems in the electricity grid. In this paper we focus on a forecast technique for the production of a photovoltaic plant, one day in advance, with the ultimate target of the optimal management of an energy storage system.

The procedure is based on a regression model that takes as input the weather forecasts of the US Global Forecasting Service (GFS) and it is trained and tested on one year of power production data of a 1.3MW plant located in Borkum, Germany. The method used is the Multilinear Adaptive Regression Splines, that allowed the automatic definition of a reasinably simple model for the system and whose regression coefficients can be easily interpreted.

The forecasted power obtained by the model proved to have a high correlation with the measured data and relatively low errors even with a limited number of features included in the model and a low number of training samples.

Keywords: photovoltaic systems; power production forecast; multilinear adaptive regression splines; numerical weather prediction

1. Introduction

Inclusion into power distribution grids of renewable energy, from photovoltaic (PV), Concentrated Solar Power (CSP) and Concentrating Photovoltaic (CPV) plants in particular and their increasing diffusion, raises a number of technological challenges that must be solved to achieve an economic profit from the integration. The variable nature of the solar resource and the difficulties introduced between balance, reliability and flexibility of the network imposes a backup for an integrated system. This can be achieved by means of auxiliary

*Corresponding author

Email addresses: luca.massidda@crs4.it (Luca Massidda), marino.marrocu@crs4.it (Marino Marrocu)

generators and/or storage systems that can intervene during periods of high variability.

Regardless of the current regulations of the electricity market in different regions, which favor, in general, the definition of the production profile one day in advance [1], [2], [3], an accurate forecast is critical to the proper dimensioning and optimal management of any storage facility for the distribution grid [4], [5] and can substantially contribute to make the integration profitable.

Several approaches have been developed in recent years, to forecast the power production obtained from renewable resources; for a recent and fairly comprehensive review see e.g. [6] and [7] and references therein. Depending on the time horizon for which the solar power forecast is needed, different approaches are, in general, required. Short-term forecast, in which the horizon is lower than one hour ($T_F < 1h$), is based mainly on sky imaging techniques or time series analysis, while the use of satellite data can contribute to the accuracy of the prediction when the forecast horizon increases ($1h < T_F < 6h$) [8] [9]. Due to the intrinsic non-linearity of the atmospheric circulation, to which variability of solar energy production is strongly tied, accurate energy forecast for longer time horizons ($T_F > 6h$) must necessarily rely upon the use of the state of the art of Numerical Weather Prediction (NWP).

In this paper we focus on a $24h$ forecast technique for the production of a photovoltaic plant, with the ultimate target of the optimal management of an energy storage system. The developed forecast system, and the results discussed here, will be based solely on information that may be obtained from the weather conditions of the numerical prediction model, and from the past data on weather conditions and power production of the PV system. Therefore any information about factors affecting efficiency of the power plant, such as for example cells temperature, shading due to dust deposition, inverter efficiency, status of wiring have been neglected. The procedure is based on a regression techniques known as Multivariate Adaptive Regression Splines. The regression model uses the NWP data, without the need of measurements of the power production of the plant in the recent history. The power production history data are solely used for the training phase of the model. Finally we used a clear sky model to mimic time variability of power production at time scales smaller (15 minutes) than that of weather forecast (3 hours) .

The method is applied to the power produced by a medium sized PV plant, located in the German island of Borkum.

The proposed method bases the power forecast mainly on the Global Horizontal Irradiation (GHI) forecast which, being strongly related to the instantaneous power obtained by a photovoltaic system, is of particular relevance among the available products of a NWP model. A huge bulk of literature on the prediction of the GHI a day in advance exists, though it is not exclusively related to the specific issue of energy production. In contrast, the literature regarding the use of GHI and other output variables obtained from a meteorological model to the forecast of energy production from photovoltaic systems is scant, presumably due to the poor accuracy of PV plants production data [10].

Recent works about the forecast of the power output of PV systems using

NWP output include [11] in which 21 PV stations in Denmark were analysed using the High Resolution Limited Area Model of the Danish Meteorological Institute. A single PV plant in Spain was studied using again a high resolution model in [12]. In [13], forecasts from the Canadian Meteorological Centre, validated against ground measurements from the United States's SURFRAD network, were used to evaluate the performance of three small PV systems; weather forecast data were processed through spatial averaging and bias removal using Kalman filtering. Input coming from the Japan Meteorological Agency weather forecast model are used in [14] and [15] to predict data of PV plants production in Japan using Support Vector Machine regression techniques (SVR). The same method has been applied to a single photovoltaic power plant in [16]. A method based on global and meso-scale weather prediction models and artificial neural network was adopted and compared with other methods in [17] for a PV plant located in Spain. The power output of five tracking plants in Spain have been recently modelled using as input the forecast of several NWP models and a non-parametric approach (the Quantile Regression Forests as machine learning tool) [18]. Recently forecast of the Ensemble Prediction System (EPS) of ECMWF and artificial neural networks were used in [19] to produce a probabilistic forecast of the power production of three solar farms located in Italy.

The paper is structured as follows. In section 2 the NWP model data sets are described, then in section 3 a brief description of the forecasting method for the power production of the plant is presented. In section 4 the regression technique is briefly described. Sections 5 and 6 contain the description of the naive persistence model used as a benchmark and the performance measures adopted. The results of the application of the methodology and evaluation of its performance are discussed in section 7. Finally in section 8 we summarise main findings of this study and outline possible future improvements that may be implemented.

2. Numerical Weather Prediction

The main purpose of this work was to verify the maximum degree of skill achievable in the prediction of PV energy production using as predictors the meteorological fields produced by a numerical weather forecast model and to quantify the benefit that can be obtained in comparison with the simplest predictive methods based on persistence and on the knowledge of the climate.

Therefore, our first objective was to separate the uncertainty introduced by the predictors being used from that related to the length of the forecast time. To achieve this result we used two different data sets for the calibration and verification of the procedure. In the first one (named from now on GFS1) we used forecast data with the shortest possible forecast time, instead for the second (GFS2) we used data with forecast time greater than $24h$.

Moreover the "ideal" weather forecast data set used in our analysis, basically needs to:

- be available for 2014, the year for which the data on production of the photovoltaic plant under study are available;
- be updated multiple times per day, so as to support a real-time operating procedure
- possibly be publicly available.

These features are fully satisfied by the global model forecasts GFS (Global Forecasting System) operated by the US National Meteorological Service[1]. As for many of the activities funded by the US, the GFS model data are made publicly available and form the basis for many commercial and research activities, including private ones. The GFS model is a spectral model operated 4 times a day, starting from the time of analysis 00, 06, 12 and 18 UMT, and provides global forecasts, with an average of about 13km horizontal resolution and 64 vertical levels. Constantly updated forecasts are available at the maximum spatial resolution, for forecasting time up to +120hr, every 3 hours (actually, since May 2016 forecast times up to +120hr are available hourly).

Past data of the operational model for the year 2014, although at reduced spatial resolution, have been downloaded from dedicated servers operated by NOAA (National Oceanic and Atmospheric Administration).

More precisely, the dataset GFS1, which we used to train the regression model, is obtained collecting the data for year 2014 at 0.5° resolution for +03hr and +06hr forecast time. This allowed reconstructing the variability of meteorological fields during the day with a temporal resolution of 3 hours. In detail, we used the forecast + 03hr and +06hr of the analysis of

- 00 UTM, for 03 and 06 UTM,
- 06 UTM, for 09 and 12 UTM,
- 12 UTM, for 15 and 18 UTM,
- 18 UTM, for 21 and 00 UTM.

To estimate the skill of the procedure when using predictors at longer forecast time, we build up the GFS2 data set, covering the second part of the year 2014, using fields for times of forecast +27hr and +30hr. The GFS2 data set has been organised similarly to GFS1, in which however the +27hr (+30hr) forecast has been used for each day in place of the +3hr (+6hr) and the analysis time is that of the day before (24hr ahead).

3. The forecasting model

The power that can be obtained from a photovoltaic plant is almost entirely related to the solar radiation incident on the panels from the ground or the sky, both as direct and diffuse solar radiation.

[1] http://www.emc.ncep.noaa.gov/index.php?branch=GFS

The incident solar radiation on top of the earth's atmosphere fluctuates around an average value estimated at $I_0 = 1360 W m^{-2}$ [20]. This is attenuated in its way to the earth's surface due to complex multiple reflections, absorptions, reemissions from various layers and it is modulated by the various components of the atmosphere as for example areosols and water vapour. The incident radiation on the earth's surface is divided in two distinct components: the Direct Normal Irradiance (DNI) and Diffuse Horizontal Irradiance (DHI) whose geometric sum is equal to the Global Horizontal Irradiance (GHI) at the ground:

$$GHI = DHI + DNI \cdot \cos\theta_Z \tag{1}$$

where θ_Z is the solar zenith angle [6].

Global Horizontal Irradiance (GHI) is the amount of terrestrial irradiance falling on a horizontal surface and is available as an output from NWP models such as the GFS. In general, instead, the direct and diffuse components are not a direct model output. Several parameterisations are available for the estimation of DNI, such as the DISC and DIRINT models developed by E. Maxwell[21] and R. Perez et al.[22] respectively. In this case, DHI can then be simply obtained from equation 1.

The radiation reaching a non horizontal surface depends on its orientation, the direct and diffuse components, the radiation reflected and scattered by the earth's surface, and finally the effect of any shading or obstacle.

The radiation that reaches the Plane Of Array (POA) is:

$$E_{poa} = E_b + E_g + E_d \tag{2}$$

where E_b is the beam component, E_g is the ground-reflected, and E_d is the sky-diffuse component. The beam component of the solar irradiance is the projection of the Direct Normal Irradiance (DNI) on the surface:

$$E_b = DNI \cdot \cos AOI \tag{3}$$

where AOI is the Angle of Incidence between the Sun's rays and the surface. E_g depends on the reflectivity of the ground surface known as *albedo* and the tilt angle θ_T:

$$E_g = GHI \cdot albedo \cdot \frac{1 - \cos\theta_T}{2} \tag{4}$$

Several models have been published using different approaches to describe the irradiance from the sky dome on a tilted surface [23]. One of the simplest is the Isotropic Sky Diffuse model, relating E_d to the Diffuse Horizonthal irradiance (DHI):

$$E_d = DHI \cdot \frac{1 + \cos\theta_T}{2} \tag{5}$$

Apart from GHI another output of the weather forecast showed a good correlation with the plant's power measure, namely the *total cloud cover* (*tcc*), indicating the cloudiness in a range of (0,1).

We generate two time series with time interval Δt_m, and we denote by $ghi(t_m)$ the average total radiation to the ground and with $tcc(t_m)$ the total cloud cover. Δt_m is the weather forecast time interval (in our case $3h$) that is generally different from that of the measured power production ($15m$).

We calculate the plane of array irradiation ($E_{poa}(t_m)$) from the orientation of the panel, the position of the sun, the albedo, using the isotropic model for the diffuse radiation and the values $ghi(t_m)$, $dni(t_m)$ and $dhi(t_m)$ calculated from the meteorological model data. We also calculate the mean value of the measured power time series on the same time intervals of the GFS model, and denote it by $p(t_m)$.

The independent variables for the regression model are E_{poa} and, tcc, and the target variable is the measured power produced by the plant p. Given these data for a significant training time interval, we train a regression model based on Multivariate Adaptive Regression Splines, to estimate the time series of the average power produced with the same time frequency of weather model $\hat{p}(t_m)$.

$$\hat{p}(t_m) = f(E_{poa}(t_m), tcc(t_m)) \tag{6}$$

The model developed allows us to make a forecast of the PV plant's production averaged over the time interval between forecasts. Our main objective, however, is to obtain a prediction model that produces results that may be directly compared with the measurement of the power, it is therefore necessary to introduce an interpolation technique to calculate a higher frequency time series.

To achieve this we use a variant of the clear sky index k_t, defined as the ratio between the measured global radiation and the corresponding clear sky value :

$$k_t = \frac{GHI_t}{GHI_{cs}} \tag{7}$$

We introduce a clear sky performance index k_{pv}, defined as the ratio of the real produced power and the power that the plant would have produced in clear sky conditions:

$$k_{pv} = \frac{p}{p_{cs}} \tag{8}$$

A model for the Clear sky insolation conditions is required by almost all forecasting models [6]. The Clear Sky radiation models are developed using one or more Radiative Transfer Models (RTM) and require the knowledge of local weather variables, such as the ozone and the water vapor content in the atmosphere, the Linke turbidity [24] and the relative location of the sun. However, a recent comparison of such parametric models [25] showed that if accurate weather information are not available, even simple models, based on standard atmosphere composition, like the mESRA [26] or the Ineichen [24] models may provide satisfactory results.

We calculate the theoretical power produced by the plant for clear sky conditions with the same time frequency of measured data ($p_cs(t_p)$) and we take the mean value of this time series for the same time intervals of the weather forecast ($p_cs(t_m)$). We use the Ineichen model for the calculation of the solar radiation

components to the ground. The plane of array irradiance (E_{poa-cs}) is then calculated, given the orientation of the plant, using a simple isotropic model for the diffuse component and assuming an $albedo = 0.25$ for the calculation of the ground reflected radiation. The clear sky power is estimated as:

$$p_{cs} = p_{peak} \frac{E_{poa-cs}}{E_0} \tag{9}$$

where p_{peak} is the peak power of the plant and $E_0 = 1000 W m^{-2}$ is a reference irradiation.

Using the measured average power, we get the averaged clear performance index:

$$k_{pv}(t_m) = \frac{\hat{p}(t_m)}{p_{cs}(t_m)} \tag{10}$$

This is then filtered from outliers, that can be generated near sunrise and sunset time when $p_{cs}(t_m)$ is close to zero, and interpolated at the same instants of the measured power time serie, to get a new set of values $k_{pv}(t_p)$.

Finally, given the clear sky theoretical power $p_{cs}(t_p)$ we can estimate the power produced by the plant, as:

$$\hat{p}(t_p) = k_{pv}(t_p)\, p_{cs}(t_p) \tag{11}$$

The value of the energy produced in a day is a quantity of interest and is calculated as the integral of the estimated power:

$$\hat{e}_{1d} = \int_{t=t_0}^{t_0+1d} \hat{p}(t)\, dt \tag{12}$$

4. The multivariate adaptive regression splines

The method adopted to obtain the regression model is the so called Multivariate Adaptive Regression Splines (MARS), introduced by Friedman in 1991 [27]. A large number of methods have been used in literature, for the forecast of renewable power production from weather variables, such as linear and non-linear regression, regression trees, neural networks, support vector machines among others [28], but to the best of our knowledge this is one of the first applications of MARS in renewable energy forecasting.

The technique has become popular in particular for the "data mining", as it does not make any assumption or sets any particular class of relationship between the input variables and the dependent variable. This allows to synthesise an accurate model even in cases in which the relationship between the variables is not monotonous or can be hardly approximated by parametric models. MARS builds a functional relationship as a set of coefficients and basis functions, solely from the available data, using a "divide and conquer" strategy: the input space is divided into regions and for each of these a regression equation is evaluated.

The equation generated by the model assumes the generic form:

$$\hat{f}(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \tag{13}$$

It is therefore dependent on the vector of variables X used as predictors and on their cross products, and consists of an intercept term β_0 and the weighted sum by the coefficients β_m of one or more basis functions.

Each basis function $h_m(X)$ can be a *hinge function*, or the product of two or more *hinge functions*. A *hinge function* has the form $max(0, x-t)$ or $max(0, t-x)$, where t is a constant called *knot*. The algorithm searchs on the whole space of possible inputs and the corresponding output values to detect and automatically select the variables to be included in the model, the number of the basis functions and the values of the *knots*. During the search, a growing number of basis function are added to the model, minimising the root mean square error between the measured and predicted output of the model. The most important independent variables and the most significant relationships between them are determined automatically. The result is not necessarily a piecewise linear function as might appear at first glance, as the basis functions of the model can also be formed by the product of *hinge* functions, giving rise to non-linear models.

Once this phase is completed, the model is refined by eliminating the basis functions that are associated with minimal increases of accuracy of the fit. This is the "Generalised Cross Validation error", which takes into account not only the residual error but also the complexity of the resulting model, reducing, at the same time, the risk of overfitting. Its expression is:

$$GCV = \frac{\sum_{i=1}^{N} (y_i - f(x_i))^2}{\left(1 - \frac{C}{N}\right)^2} \tag{14}$$

where N is the number of samples of the data set and x_i and y_i are the input and output variable. $C = 1 + cd$, with d equal to the number of independent basis functions and c is a penalty parameter between two and three.

5. Persistence model

To evaluate the performance of the forecasting procedure adopted we use as benchmark the persistence model. This approach exploits the autocorrelation property of a time series and, as the name suggests, it is based on the assumption that the same conditions persist over time or occur with a known time interval.

$$p(t_m) = p(t_m - \Delta t) \quad and \quad p(t_p) = p(t_p - \Delta t) \tag{15}$$

In our case, in order to get a prediction of energy production on the following day we will assess the performance of the model with $\Delta t = 1day$.

As for the power, the total energy produced during the day will be calculated as:

$$e_{1d}(t) = e_{1d}(t - \Delta t) \tag{16}$$

6. Performance measure

The performances of a solar plant forecasting procedure are evaluated using several indicators in the current literature, often making it difficult to compare the results.

Some parameters are used quite extensively, such as the coefficient of determination R^2, the Root Mean Square Error (RMSE) , the Mean Average Error (MAE) and the Mean Bias Error (MBE).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{17}$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(x_i - \hat{x}_i)^2}{n}} \tag{18}$$

$$MAE = \sum_{i=1}^{n}\frac{|x_i - \hat{x}_i|}{n} \tag{19}$$

$$MBE = \sum_{i=1}^{n}\frac{x_i - \hat{x}_i}{n} \tag{20}$$

where x_i are the measured data, $\hat{x}_i$ are the forecasts and $\bar{x}$ is the mean value of the measured data.

The algorithm's performance will be evaluated as the ability to forecast the instantaneous power compared with the measured values with a $15m$ cadence. The performance will also be assessed on a time series with an average time step of $3h$, at the time instants where the time series of the GFS meteorological forecast data are given. When calculating the performance of the algorithm only the daytime values are considered.

These accuracy measures are completed with the error in the estimation of the daily energy production.

7. Results

The forecast procedure is tested on a medium sized PV plant located in the German island of Borkum, south oriented, with a tilt angle of 38° and an approximate peak power of $1.3MW$. The data of the power produced by the plant is available for the year 2014 as a complete time series with a time interval of $15m$.

Since only one year of measures was available we used the first half of the year for the training of the model and the second half for the testing. For the training we used the GFS1 dataset, whose data is relative to a short forecast time, the resulting model is tested in the second half of the year on both the GFS1 and GFS2, the latter is obtained for the second half of the year with a forecast time of one day.

Table 1: MARS model fitted GFS1 data for the first half of the year, the model is calculated as a weighted sum of the basis function on the left column, with weight coefficients listed in the right columns. $h(x-k)$ is the hinge function for the variable x and knot k. The last column is the explained variance of the first term and the two last terms of the function.

Basis Function	Coefficient	Expl. var.
$(Intercept)$	1.365	-
E_{poa}	0.753	94%
$h(tcc - 0.56) \cdot E_{poa}$	0.45	6%
$h(0.56 - tcc) \cdot E_{poa}$	0.44	

The MARS method allows to select the complexity of the model obtainable. We discuss here a simple model in which we limited the input parameters to only two values, namely the solar radiation on the plane of array (E_{poa}) and the total cloud cover (tcc) and we fixed the maximum degree of the interpolating function to six. The results of the automatic model generation are presented in Table 1, showing a strong linear correlation between the power produced and the radiation on the plane of array, modulated by the forecasted total cloud cover. The model obtained has a very low value for the intercept, and only a linear term in E_{poa} and a second degree basis function are kept by the fitting procedure, with a hinge function having a knot at 56% of the cloud cover. Although 95% of the variance is described by the E_{poa} linear term alone, the addition of the tcc term, appearing by means of the second degree basis function, improves the described variance of the 5%. More complex formulations are obtainable including more input variables in the model or increasing the maximum allowed degree, but the resulting performances were similar (or worse, in some case) to the simple model shown here, with a less stable and interpretable function. This is particularly true for our operational configuration where calibration data are limited in time and not homogeneous (+3h, +6h) with those of the forecasting stage (+27h, +30h).

The accuracies of the MARS based procedure for the GFS1 and GFS2 datasets, as well as that of the persistence model, are listed in Table 2 for the $3h$ and $15m$ power time series and for the daily energy production.

The results for the $3h$ forecast are substantially more accurate that those of the persistence presenting a much higher correlation value (0.834 against 0.488) and RMSE and MAE almost halved. The low value of R^2 for the persistence model is an indicator of the variability of the local weather conditions, that are only weakly similar from day to day. The high correlation value for the MARS model is due to the high correlation of the weather variables selected with the power output of the plant. In this case we are also taking advantage of the smoothing in the measured power data due to the averaging in the $3h$ time interval. The results of the $15m$ interpolation are obviously less precise, the R^2 of the persistence model is almost halved compared to the $3h$ results. A decrease could be noticed also for the MARS regression, since we are lacking detail in

Table 2: Performance comparison for the forecast methods for power with $3h$ and $15m$ time cadence and for daily energy production. Power values are in kW, energy is measured in kWh, except for the R^2 coefficient.

Variable	Mean	Peak	Method	R^2	RMSE	MAE	MBE
p_{3h}	263.8	1064.4	Persistence	0.488	235.1	146.3	2.4
			MARS GFS1	0.834	119.2	78.6	3.0
			MARS GFS2	0.805	128.7	86.5	3.5
p_{15m}	332.3	1326.0	Persistence	0.262	286.2	189.2	-0.5
			MARS GFS1	0.735	168.8	117.8	4.4
			MARS GFS2	0.706	177.7	125.9	5.0
e_{1d}	4060.8	10047.8	Persistence	0.452	2076.1	1523.6	1.8
			MARS GFS1	0.868	1012.7	769.9	50.3
			MARS GFS2	0.843	1106.1	862.5	57.8

weather information and only a smoothed variation of the power production profile can be obtained. This is even more evident in the other forecast error measures, where the limits of the interpolation procedure and the lack of higher frequency weather forecasts appear.

As expected the performance of the GFS2 dataset is worse than that obtained for the GFS1 data due to the longer forecast horizon, but the decrease of the performance measures is relatively small. This testifies the accuracy of the forecast provided by the GFS, since increase in the forecast horizon has a limited effect on the accuracy of the power and energy from the plant. The MAE is in fact equal to 8.9% of the peak power and to 35.4% of the mean power for the GFS1 dataset and equals 9.5% of the peak power and 37.9% of the mean power for the GFS2 data. Looking at the energy production the MAE is 7.7% of the peak and 18.9% of the mean power for GFS1 and 8.6% and 21.2% of the peak and mean power for the GFS2.

The higher accuracy of the GFS1 dataset is also noticeable in the two figures 1 and 2 in which the real data (measure), the forecast with the shortest forecast times (GFS1) and the forecast with longest forecast time (GFS2) are shown, for a relatively sunny week and a more cloudy week at the end of 2014.

Long periods of clear sky conditions with associated maximum photovoltaic production are not common, in Borkum. This is particularly evident from the image on the left in figure 3, showing the power production during all the year, and where it is evident that the production is characterised by strong oscillation at a relatively high frequency.

The low frequency at which the weather forecast data are available cannot reproduce such variability, and the image obtained from the model (shown on the right) appears as a smoothed version of the real one, while still being able to reproduce its main patterns.

We expect to have better performance in the near future using the same

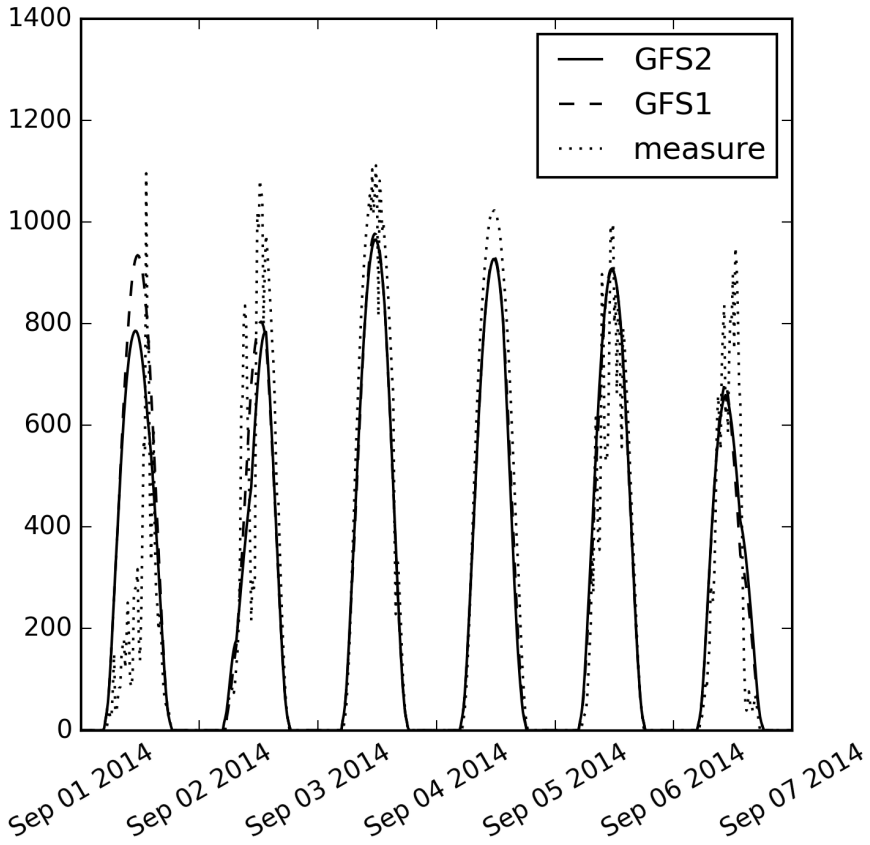

Figure 1: Measured and forecast power for a sunny week

method since the GFS system has recently revised its forecast output increasing its temporal resolution from $3h$ to $1h$ that will permit the detection of higher frequencies of the power output signal.

8. Summary and conclusions

In this paper a new procedure to forecast the power production of a photovoltaic power plant, 24h in advance, is described. The regression model developed uses past historical plant power output data and the publicly available output data of the weather forecasts of the GFS numerical weather prediction model.

The model is based on Multilinear Adaptive Regression Splines, a method that, to our knowledge, has not yet been applied to renewable energy prediction

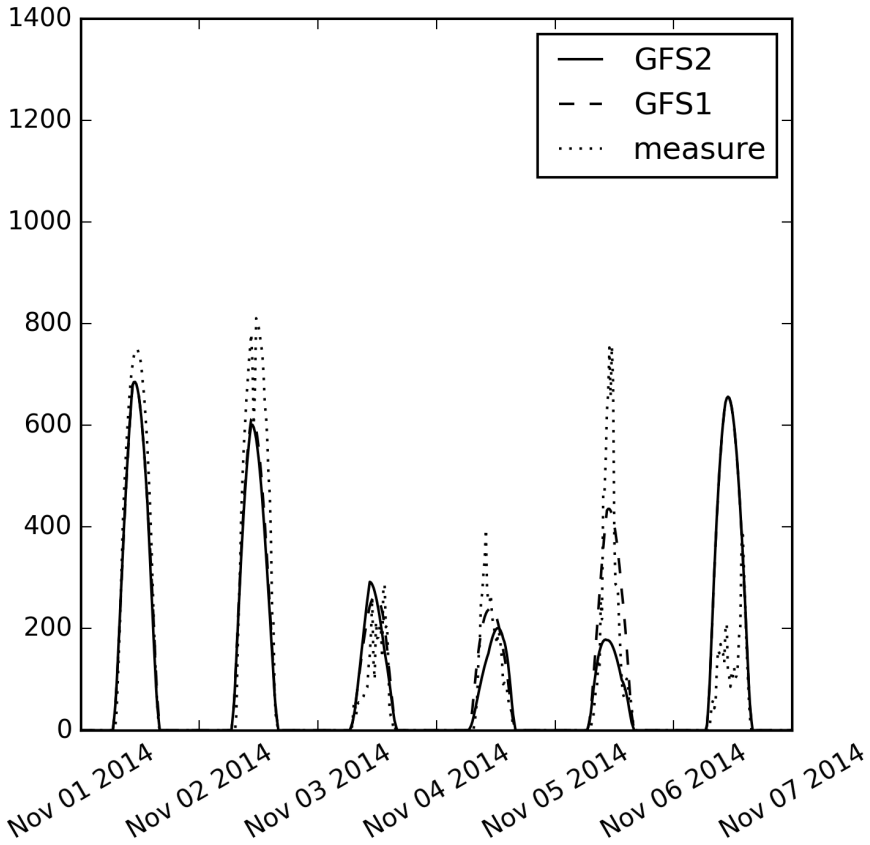

Figure 2: Measured and forecast power during a cloudy week

and that allows to obtain a non linear but still interpretable regression function from a set of well chosen input variables.

The output data from the GFS model with the highest correlation with the PV power production have been individuated and, on these, the regression model have been fitted, giving the mean power output on the $3h$ time intervals that are characteristic of the weather forecast. Using an interpolation procedure based on a clear sky model we obtained the power forecast with the temporal resolution of $15min$, identical to that of the measures.

The procedure has been applied to a medium sized ($1.3MWp$) PV plant located in the island of Borkum (Germany) in the framework of the European H2020 project NETfficient, with one year of production data available.

Considering the relative low number of samples and features taken into ac-

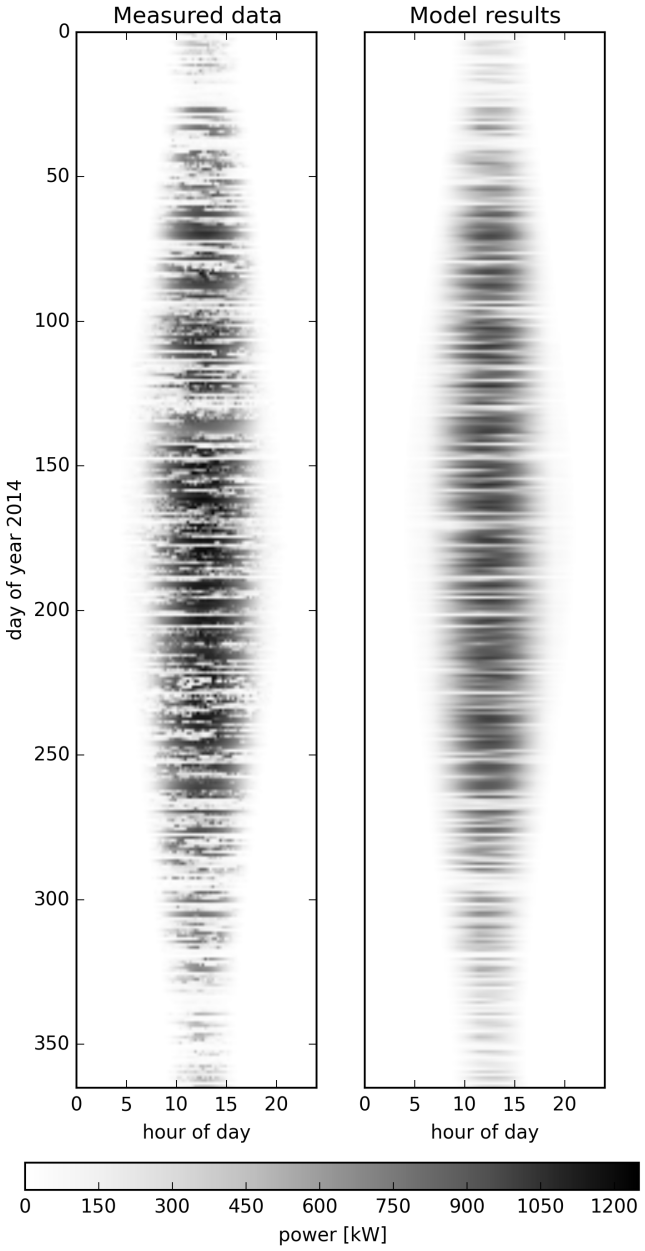

Figure 3: Measured and modeled power during the year

count, the results are promising. The forecast errors with one day horizon, are $RMSE = 177.7kW$ and $MAE = 125.9kW$ equivalent to 13.4% and 9.5% of the plant's peak power and to 53.4% and 37.9% of the mean annual power.

The errors of daily energy production of the plant, are $RMSE = 1106.1kWh$ and $MAE = 862.5kWh$ equivalent to 11.0% and 8.6% of the plant's maximum daily energy production, and to 27.2% and 21.2% of the mean.

The method is subject to further improvements, for example due to the recently improved frequency of the GFS forecasts, or by including a higher number of input variables and allowing a more complex regression function; an extended set of measured data can improve the accuracy of the regression due to a more detailed statistics and, finally, the availability of real time measurements of power production could allow the implementation of a continuous update of the calibration, resulting, hopefully, in a global improvement of forecast quality.

9. Acknowledgements

This work has been partially financed by the European Union's Horizon 2020 research and innovation programme under grant agreement No 646463, project NETfficient, Energy and economic efficiency for today's smart communities through integrated multi storage technologies.

References

[1] B. Kraas, M. Schroedter-Homscheidt, R. Madlener, Economic merits of a state-of-the-art concentrating solar power forecasting system for participation in the Spanish electricity market, Solar Energy 93 (2013) 244–255. doi:10.1016/j.solener.2013.04.012.
URL http://dx.doi.org/10.1016/j.solener.2013.04.012

[2] J. Luoma, P. Mathiesen, J. Kleissl, Forecast value considering energy pricing in California, Applied Energy 125 (2014) 230–237. doi:10.1016/j.apenergy.2014.03.061.
URL http://dx.doi.org/10.1016/j.apenergy.2014.03.061

[3] L. Nonnenmacher, A. Kaur, C. F. Coimbra, Day-ahead resource forecasting for concentrated solar power integration, Renewable Energy 86 (2016) 866–876. doi:10.1016/j.renene.2015.08.068.
URL http://dx.doi.org/10.1016/j.renene.2015.08.068

[4] R. Hanna, J. Kleissl, A. Nottrott, M. Ferry, Energy dispatch schedule optimization for demand charge reduction using a photovoltaic-battery storage system with solar forecasting, Solar Energy 103 (2014) 269–287. doi:10.1016/j.solener.2014.02.020.
URL http://dx.doi.org/10.1016/j.solener.2014.02.020

[5] M. Delfanti, D. Falabretti, M. Merlo, Energy storage for PV power plant dispatching, Renewable Energy 80 (2015) 61–72. doi:10.1016/j.renene.2015.01.047.
URL http://dx.doi.org/10.1016/j.renene.2015.01.047

[6] R. H. Inman, H. T. Pedro, C. F. Coimbra, Solar forecasting methods for renewable energy integration, Progress in Energy and Combustion Science 39 (6) (2013) 535–576. doi:10.1016/j.pecs.2013.06.002.
URL http://dx.doi.org/10.1016/j.pecs.2013.06.002

[7] J. Kleissl, Solar energy forecasting and resource assessment, Academic Press, 2013.

[8] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker, D. Renné, T. E. Hoff, Validation of short and medium term operational solar radiation forecasts in the US, Solar Energy 84 (12) (2010) 2161–2172. doi:10.1016/j.solener.2010.08.014.
URL http://dx.doi.org/10.1016/j.solener.2010.08.014

[9] A. Zagouras, H. T. Pedro, C. F. Coimbra, On the role of lagged exogenous variables and spatio–temporal correlations in improving the accuracy of solar forecasting methods, Renewable Energy 78 (2015) 203–218. doi:10.1016/j.renene.2014.12.071.
URL http://dx.doi.org/10.1016/j.renene.2014.12.071

[10] D. P. Larson, L. Nonnenmacher, C. F. Coimbra, Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest, Renewable Energy 91 (2016) 11–20. doi:10.1016/j.renene.2016.01.039.
URL http://dx.doi.org/10.1016/j.renene.2016.01.039

[11] P. Bacher, H. Madsen, H. A. Nielsen, Online short-term solar power forecasting, Solar Energy 83 (10) (2009) 1772–1783. doi:10.1016/j.solener.2009.05.016.
URL http://dx.doi.org/10.1016/j.solener.2009.05.016

[12] D. Masa-Bote, M. Castillo-Cagigal, E. Matallanas, E. Caamaño-Martín, A. Gutiérrez, F. Monasterio-Huelín, J. Jiménez-Leube, Improving photovoltaics grid integration through short time forecasting and self-consumption, Applied Energy 125 (2014) 103–113. doi:10.1016/j.apenergy.2014.03.045.
URL http://dx.doi.org/10.1016/j.apenergy.2014.03.045

[13] S. Pelland, G. Galanis, G. Kallos, Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model, Prog. Photovolt: Res. Appl. 21 (3) (2011) 284–296. doi:10.1002/pip.1180.
URL http://dx.doi.org/10.1002/pip.1180

[14] J. G. da Silva Fonseca Junior, T. Oozeki, H. Ohtake, K. ichi Shimose, T. Takashima, K. Ogimoto, Regional forecasts and smoothing effect of photovoltaic power generation in Japan: An approach with principal component analysis, Renewable Energy 68 (2014) 403–413. doi:10.1016/j.renene.2014.02.018.
URL http://dx.doi.org/10.1016/j.renene.2014.02.018

[15] J. G. da Silva Fonseca Junior, T. Oozeki, H. Ohtake, T. Takashima, K. Ogimoto, Regional forecasts of photovoltaic power generation according to different data availability scenarios: a study of four methods, Prog. Photovolt: Res. Appl. 23 (10) (2014) 1203–1218. doi:10.1002/pip.2528.
URL http://dx.doi.org/10.1002/pip.2528

[16] J. G. da Silva Fonseca, T. Oozeki, T. Takashima, G. Koshimizu, Y. Uchida, K. Ogimoto, Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu Japan, Prog. Photovolt: Res. Appl. 20 (7) (2011) 874–882. doi:10.1002/pip.1152.
URL http://dx.doi.org/10.1002/pip.1152

[17] L. A. Fernandez-Jimenez, A. Muñoz-Jimenez, A. Falces, M. Mendoza-Villena, E. Garcia-Garrido, P. M. Lara-Santillan, E. Zorzano-Alba, P. J. Zorzano-Santamaria, Short-term power forecasting system for photovoltaic plants, Renewable Energy 44 (2012) 311–317. doi:10.1016/j.renene.2012.01.108.
URL http://dx.doi.org/10.1016/j.renene.2012.01.108

[18] M. P. Almeida, O. Perpiñán, L. Narvarte, PV power forecast using a non-parametric PV model, Solar Energy 115 (2015) 354–368. doi:10.1016/j.solener.2015.03.006.
URL http://dx.doi.org/10.1016/j.solener.2015.03.006

[19] S. Sperati, S. Alessandrini, L. D. Monache, An application of the ECMWF Ensemble Prediction System for short-term solar power forecasting, Solar Energy 133 (2016) 437–450. doi:10.1016/j.solener.2016.04.016.
URL http://dx.doi.org/10.1016/j.solener.2016.04.016

[20] G. Kopp, J. L. Lean, A new lower value of total solar irradiance: Evidence and climate significance, Geophys. Res. Lett. 38 (1) (2011) n/a–n/a. doi:10.1029/2010gl045777.
URL http://dx.doi.org/10.1029/2010gl045777

[21] E. L. Maxwell, A quasi-physical model for converting hourly global horizontal to direct normal insolation, Tech. rep., Solar Energy Research Inst., Golden, CO (USA) (1987).

[22] P. Ineichen, R. Perez, R. Seal, E. Maxwell, A. Zalenka, Dynamic global-to-direct irradiance conversion models, Ashrae Transactions 98 (1) (1992) 354–369.

[23] P. Loutzenhiser, H. Manz, C. Felsmann, P. Strachan, T. Frank, G. Maxwell, Empirical validation of models to compute solar irradiance on inclined surfaces for building energy simulation, Solar Energy 81 (2) (2007) 254–267. doi:10.1016/j.solener.2006.03.009.
URL http://dx.doi.org/10.1016/j.solener.2006.03.009

[24] P. Ineichen, R. Perez, A new airmass independent formulation for the Linke turbidity coefficient, Solar Energy 73 (3) (2002) 151–157. doi:10.1016/s0038-092x(02)00045-2.
URL http://dx.doi.org/10.1016/s0038-092x(02)00045-2

[25] P. Ineichen, Comparison of eight clear sky broadband models against 16 independent data banks, Solar Energy 80 (4) (2006) 468–478. doi:10.1016/j.solener.2005.04.018.
URL http://dx.doi.org/10.1016/j.solener.2005.04.018

[26] C. Rigollier, O. Bauer, L. Wald, On the clear sky model of the ESRA — European Solar Radiation Atlas — with respect to the heliosat method, Solar Energy 68 (1) (2000) 33–48. doi:10.1016/s0038-092x(99)00055-9.
URL http://dx.doi.org/10.1016/s0038-092x(99)00055-9

[27] J. H. Friedman, Multivariate Adaptive Regression Splines, Ann. Statist. 19 (1) (1991) 1–67. doi:10.1214/aos/1176347963.
URL http://dx.doi.org/10.1214/aos/1176347963

[28] T. Hastie, J. Friedman, R. Tibshirani, The Elements of Statistical Learning, Springer New York, 2001. doi:10.1007/978-0-387-21606-5.
URL http://dx.doi.org/10.1007/978-0-387-21606-5