Machine Learning Prediction of Large Area Photovoltaic Energy Production

Ángela Fernández, Yvonne Gala, José R. Dorronsoro

Dpto. Ing. Informática, Universidad Autónoma de Madrid

Abstract. In this work we first explore the use of Support Vector Regression to forecast day-ahead daily and 3-hourly aggregated photovoltaic (PV) energy production on Spain using as inputs Numerical Weather Prediction forecasts of global horizontal radiation and total cloud cover. We then introduce an empirical "clear sky" PV energy curve that we use to disaggregate these predictions into hourly day-ahead PV forecasts. Finally, we use Ridge Regression to refine these day-ahead forecasts to obtain same-day hourly PV production updates that for a given hour h use PV energy readings up to that hour to derive updated PV forecasts for hours $h+1, h+2, \ldots$. While simple from a Machine Learning point of view, these methods yield encouraging first results and also suggest ways to further improve them.

Keywords: Photovoltaic energy, Numerical Weather Prediction, Support Vector Regression, Ridge Regression.

1 Introduction

The constant integration of renewable energy, particularly wind and solar, and their increasing effect in the electrical systems of countries such as the USA, Germany, Denmark or Spain has as a consequence a growing needed of accurate forecasts. Usually these forecasts are needed for the day-ahead level, i.e., about 24 hours in advance, or for same-day hourly updates of previous forecasts that are given for the next few hours. In the case of Spain, 24 hour day-ahead forecast are required for daily energy markets and same-day 4 to 8 hour updates for intra-day markets. A large forecast effort has been carried out in the past years for wind energy and has resulted in a wide use of Machine Learning (ML) models and tools such as Neural Networks (NN) or Support Vector Regression (SVR) to predict either local energy production at single wind farms or global energy production over a wide geographic area. These NN or SVR models have as their inputs the day ahead Numerical Weather Prediction (NWP) forecasts provided by systems such as the Global Forecasting System (GFS, [5]) or the European Center for Medium Weather Forecast (ECMWF, [4]) and give very good results for day ahead prediction. On the other hand, short term wind energy forecasting is more difficult, as simple persistence prediction is very hard to beat at the first 1–2 hours, simple models are usually competitive for the next few hours and any

model reverts to the NWP-based day-ahead forecast in about 5–8 hours at single wind farms and 10–14 hours for very wide area predictions. We point out that wind farms are usually not operated, in the sense that wind energy is directly converted into electric energy without the farm's output being regulated. Thus, the farm works on a stable basis and model building and prediction can thus be carried through in a time homogeneous basis.

Solar energy forecasting is following basically the steps already taken for wind energy. There are two main solar technologies, thermosolar or concentrating solar power (CSP), where the Sun's radiation is used to heat up a fluid that then drives a steam turbine to produce electricity, and photovoltaic (PV), where radiation is transformed directly into electricity. CSP plants are usually operated to a high degree, particularly if they have facilities such as molten salt tanks to store heat that can be used to generate energy after dark. This makes obviously quite difficult their direct energy forecasting by, say, ML methods. On the other hand, PV plants are at this moment not so operable and, thus, are in principle more amenable to direct ML energy forecasts. However, they are relatively small and, thus, very sensitive to cloud effects. Moreover, PV energy production has no inertia, which results in very sharp and wide fluctuations which as of today, are essentially impossible to predict for single plants. A further difficulty is the fact that solar NWP forecasting is much less developed than wind NWP; in fact, wind energy has been a driving force in the past few years for the NWP systems to adapt to the industry needs. However, this is not yet the case with radiation, cloud or aerosol modeling.

In any case, the efficiency of PV cells is constantly improving and the costs of installation and operation of roof top micro PV plants are falling. This will likely lead to a potentially large increase of small and decentralized but still grid-connected micro plants. To some extent this is already the case in Spain, where there are about 4 GW of installed PV power but perhaps less than 1 GW corresponds to "large" (i.e., above 2 MW) PV plants. The local prediction of their output will be probably a too challenging problem in the near future; on the other hand, the accurate forecasting of the PV output of a much wider area should be more manageable.

In this paper we consider a particular and very large case, the hourly prediction of the global PV energy over peninsular Spain. We will consider both the day-ahead scenario, where energy predictions for the 24 hours of day d are given some time at day $d-1$, and the same-day (or short-term) prediction updates at a given hour h of previous day-ahead forecasts for hours $h+1, h+2, \ldots$. We shall use as model inputs in this first case the ECMWF NWP predictions of two meteorological variables, the aggregated downward surface solar radiation (DSSR) and the average total cloud cover (TCC), given every three UTC hours (i.e., 8 values per day) over the 1,128 points of a grid that covers essentially the entirety of the Iberian peninsula. More precisely, the DSSR values at hour h contain the sum of radiation at hours $h-2, h-1, h$; a three hour TCC value gives the sky fraction covered at that time by clouds.

There is a growing number of publications on statistical and ML modeling of radiation (certainly the most relevant variable for PV energy) but also of PV energy itself. Two recent examples of such a use of ML methods are the benchmark performed under the Weather Intelligence for Renewable Energies WIRE COST Action[1], and the 2013 competition jointly organized by the American Meteorological Society and the company Kaggle[2]. In the first case, Quantile Regression was used to derive PV estimates under a clear sky assumption. In the second, the goal was to predict daily aggregated radiation at a number of weather stations in Oklahoma from an ensemble set of NWP predictions; the winning model had as its basis Gradient Boosting Regression. On the other hand, NNs are applied in [1] to predict daily average radiation values. With respect to PV prediction, [9] reviews several Artificial Intelligence techniques in PV energy and [11,12] consider a number of ML methods. A good general reference for the many issues present in radiation and PV energy is the recent book [7]. We also point out to the 2013 Data Analytics for Renewable Energy workshop, where [14] gives a thorough overview of the modeling process of solar energy, discussing, among others, several ML approaches, and where short term PV energy forecasting is studied in [15] for both single—and aggegated—PV plants in Germany, with SVR being one of the methods considered.

In this work we will also use SVR [13] for day-ahead prediction and Ridge Regression (RR) [6] for same-day hourly forecast updates. A first reason for the SVR choice is that it is one of the most successful and widely used approaches for non linear modeling. An obvious alternative are Multilayer Perceptrons, but the large dimension of the studied problems implies that, if used, some dimensionality reduction has to be applied to input patterns so that network complexity is manageable. Another reason is the availability of the top quality software LIBSVM package [2]. Linear models are probably too simple for the fist task but, on the other hand, they are well suited to same-day predictions, given the simplicity of the delay vectors used then as model inputs.

The main goal here is obviously to obtain accurate hourly PV energy predictions. However, this cannot be done directly from NWP forecasts, as they are given only every 3 hours. Because of this, our first goal in day-ahead prediction will be to forecast three-hour aggregated energy at UTC hours $h = 3k$, $k = 0, \ldots, 7$; of course, there is no aggregated PV energy at hours 0 and 3, but in Spain there is aggregated energy at hours 6 and 21 from mid spring to mid fall. Given the 1,128 NWP grid used, pattern dimension is thus $2,256 = 2 \times 1,128$. A second goal will be to predict the aggregated daily PV energy over Spain. Here we use the six NWP predictions at UTC hours 6 to 21 as the input patterns, with now a rather large dimension of $13,536 = 6 \times 2,256$. In any case, we use these 3-hour or daily forecasts as first step towards our main goal of obtaining hourly PV forecasts, which we do by disaggregating the 3-hour and daily predictions. A first idea would be to interpolate them using some physical clear sky radiation

[1] `http://wire1002.ch/fileadmin/user_upload/Documents/ES1002_Benchmark_announcement_v6.pdf`.

[2] `https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest`.

model. A review of some of these models from the point of view of renewable energy is in [10]. However, clear sky models have a very local nature while here we have to disaggregate energy produced all over peninsular Spain and there is no simple way to define a physical model for such a large area. We follow another approach, working with an "empirical clear sky" energy model that we build taking as a basis the hourly maximum normalized PV energy values observed over a number of years. Our results give Mean Absolute Errors (MAE) for the day light hours of about 2–3% of installed power, with MAE peaks of about 5% at mid day. While root mean square errors are also widely used, we prefer to work with MAE values as they can be directly translated to energy deviations and, hence, are widely used in the industry. Of course, our errors must be compared with error values obtained elsewhere in the literature but it seems that more research effort has been devoted to short-term PV energy forecasting than to the day-ahead problem and, thus, day-ahead error reference values seem hard to come by.

For same-day PV energy prediction updates, the hourly readings of the PV energy time series is basically the only information available. Because of this and denoting by $\mathcal{E}_{dk}$ the PV energy reading at hour k of a day d, we will simply use a RR linear model where for a given day d, a delay vector $(\mathcal{E}_{d6}, \ldots, \mathcal{E}_{dh-1}, \mathcal{E}_{dh})$ is built at hour h to predict PV energy values $\hat{\mathcal{E}}_{dk}$ at hours $k = h+1, \ldots, 20$. The choice of the 6 to 20 UTC hour interval corresponds to day-light hours in Spain for most of the year.

The paper is structured as follows. We very briefly review SVR and RR in Sect. 2. In Sect. 3 we will describe our 3-hour and daily aggregated PV models and numerically compare their behavior. The empirical clear sky PV energy model is described in Sect. 4 and applied to disaggregate to the hourly level the predictions given by the best 3-hour and daily models; we will also give error comparisons. In Sect. 5 we discuss the RR-based same-day models and the paper ends with a brief discussion and conclusions section.

2 Ridge and Support Vector Regression

Assuming a $N \times d$ dimensional data matrix $\mathcal{X}$ associated to N patterns X_t with dimension d, in linear regression we want a weight vector W such that $X_t \cdot W \simeq y_t$, $t = 1, \ldots, N$, or, in matrix notation, $\mathcal{X}W \simeq Y$, where we recall that the rows of $\mathcal{X}$ contain the transposes of the X_t patterns and Y is the N-dimensional output vector with y_t as components. For simplicity we assume the X_t, y_t to have zero mean. In RR the optimal weight W^* is found by minimizing the error function

$$\frac{1}{2}\frac{1}{N}\|\mathcal{X}W - Y\|_2^2 + \frac{\lambda}{2}\|W\|_2^2.$$

In the optimal W^* we have thus a balance between model error and complexity, which we control through the parameter λ. W^* can be found analytically as

$$W^* = (\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\mathcal{X}^T Y,$$

where I is the $d \times d$ identity matrix. Since $\mathcal{X}^T\mathcal{X}$ is positive semidefinite, $\mathcal{X}^T\mathcal{X}+\lambda I$ is invertible for any $\lambda > 0$. We have to decide on the optimal value of λ, which is usually determined by some form of Cross Validation (CV). While not very powerful to tackle general complex problems, the simplicity of RR makes it suitable for relatively simple problems; we will use it here for same-day, hourly PV energy prediction updates.

The cost function of RR can be written as

$$\sum_t \ell(y_t, W \cdot X_t) + \frac{1}{C}\|W\|^2$$

with $\ell(y,z) = \frac{1}{2}(y-z)^2$ the quadratic loss and $C = \frac{1}{N\lambda}$. Assuming now a non-homogeneous model $W \cdot X + b$, the cost function in SVR [13] is

$$\sum_t [y_t - W \cdot X_t - b]_\epsilon + \frac{1}{C}\|W\|^2 \tag{1}$$

with the loss now being $\ell(y,z) = [y-z]_\epsilon$ and $[v]_\epsilon = \max\{|v| - \epsilon, 0\}$ the ϵ-insensitive loss. We thus allow an ϵ-wide, penalty-free "error tube" around the model. To solve (1), it is rewritten as a constrained minimization problem:

$$\min_{W,b,\xi} \frac{1}{2}\|W\|^2 + C\sum_t (\xi_t + \xi_t^*), \tag{2}$$

subject to the restrictions $W \cdot X_t + b - y_t \geq -\xi_t - \epsilon$, $W \cdot X_t + b - y_t \leq \xi_t^* + \epsilon$ and $\xi_t, \xi_t^* \geq 0$. Its dual problem is then obtained by Lagrangian theory, that yields

$$\begin{aligned} \min \Theta(\alpha,\beta) = & \frac{1}{2}\sum_{t,s}(\alpha_t - \beta_t)(\alpha_s - \beta_s) X_t \cdot X_s + \\ & \epsilon\sum_t(\alpha_t + \beta_t) - \sum_t y_t(\alpha_t - \beta_t) \end{aligned} \tag{3}$$

which has much simpler box constraints $0 \leq \alpha_t, \beta_t \leq C$. The Karush–Kuhn–Tucker (KKT) conditions for problems (2) and (3) can be applied to compute the optimal W^*, b^* of the primal problem from the optimal dual solutions α_t^*, β_t^*, yielding a final model

$$f(X) = f(X, W^*, b^*) = W^* \cdot X + b^* = \sum(\alpha_t^* - \beta_t^*) X_t \cdot X + b^*.$$

Notice that $f(X, W^*, b^*)$ is also a simple and perhaps not powerful enough linear model. Since f and (3) only involve dot products, the kernel trick [13] can be used to build f not on the original inputs X but on their extensions $\phi(X)$ to a possibly infinite dimensional Hilbert space. To do so we use an appropriate kernel $K(X, X')$ so that we have $\phi(X) \cdot \phi(X') = K(X, X')$. Thus, we can compute dot products on the extended $\phi(X)$ patterns without actually having to handle them explicitly. Using a Gaussian kernel $e^{-\frac{\|X - X'\|^2}{2\sigma^2}}$ results in a final model

$$f(X) = b^* + \sum_t (\alpha_t^* - \beta_t^*) e^{-\frac{\|X - X_t\|^2}{2\sigma^2}}.$$

Finally, we have to select three parameters, C, ϵ and the Gaussian kernel width σ, which we do again by CV, much costlier now than for RR. In our experiments we will use LIBSVM [2] for SVR and a Matlab's RR implementation.

3 Daily and 3-hour Wide Area PV Energy Forecasting

We discuss next the use of SVR to derive daily and 3-hour forecasts of aggregated PV energy. Our data sources are hourly global PV energy values in peninsular Spain from June 1, 2011 to May 31 2014, and NWP forecasts of DSSR and TCC from the ECMWF from December 2012 to May 2014. The NWP period is much smaller, so for testing we will use the six months from December 2013 to May 2014. There is a very small number of missing values in the PV energy records that, nevertheless, is not significant. In particular, there are no missing values in the test period. The input for the 3-hour aggregated energy predictions are the ECMWF forecasts of DSSR and TCC; we recall that pattern dimension is $2,256$. NWP values are given as 3-hour accumulated values for UTC hours 0, 3, 6, 9, 12, 15, 18 and 21. Hours 0 and 3 correspond in Spain to night-time all year long and, thus, we will disregard them. Three-hour PV energy values at hours 6 and 21 are also very small or zero from mid fall to mid spring but not so for the rest of the year and we will keep them. Therefore, we will first predict 3-hour accumulated energy for UTC hours 6, 9, 12, 15, 18 and 21 from the corresponding ECMWF forecasts of DSSR and TCC.

We will consider two SVR models. The first one is a single yearly model built using NWP and PV energy values from December 2012 to November 2013 and that we will test in the period from December 2013 to May 2014. We will refer to this as the yearly model. The second approach will be to build month-specific models where the model for month m is built using data from the previous $m-14$, $m-13$, $m-12$, $m-11$, $m-10$ months. For instance, the model for May 2014 is built using data from March, April, May, June and July 2013. We will refer to these as the 5-month models.

PV energy varies greatly along the day, possibly because of the different transfer behavior of PV cells. Because of this, we will build the yearly and 5-month models using submodels targeted to specific hours. For the yearly models we will use three different SVR sub-models tailored to UTC hours 6 and 21 (sunrise and sunset), 9 and 18 (mid half day) and 12 and 15 (noon). The 5-month models will be built upon two different SVR sub-models, one of them tailored to hours 6 and 21 and the other to the remaining 9, 12, 15 and 18 hours; we do this as there are fewer training patterns for them. In our experience this approach yields better results than those obtained using single, all-hour SVR models.

We denote the target aggregated production values for day d and 3-hour h as E_{dh} and their corresponding predictions as $\hat{E}_{dh}^{3;y}$ for the yearly model and $\hat{E}_{dh}^{3;5m}$ for the 5-month model. While there are actually three yearly and two 5-month underlying models, we will report their joint results as those of single models. We denote their errors for day d and 3-hour h as $e_{dh}^{3;y}$ and $e_{dh}^{3;5m}$ respectively.

We will also build similar models to predict now the aggregated daily PV energy generated in Spain. Recall that pattern dimension will then be $13,536 = 6 \times 2,256$, as there are six NWP forecasts for each 3-hour interval from hours 6 to 21. Again we will consider a yearly model built using the entire year from December 2012 to November 2013 and also the 5-month models described above. Here we denote the target aggregated production for a day d as E_d, their corresponding predictions as $\hat{E}_d^{D;y}$ for the yearly model and $\hat{E}_d^{D;5m}$ for the 5-month models, and their errors as $e_d^{D;y}$ and $e_d^{D;5m}$ respectively.

An obvious alternative to these direct daily models is to predict the aggregated daily PV energy just as the sum of the 3-hour predictions; we denote now as $\hat{E}_d^{3-D;y}$ the predictions obtained using a yearly model and $\hat{E}_d^{3-D;5m}$ those using the 5-month models. Similarly, we could also consider 3-hour PV energy predictions obtained disaggregating the daily predictions $\hat{E}_d^{D;y}$ and $\hat{E}_d^{D;5m}$ into 3-hour predictions $\hat{E}_{dh}^{D-3;y}$ and $\hat{E}_{dh}^{D-3;5m}$. To do so we need for each day d a sequence $\gamma_h^d, 0 \leq h \leq 23$, with $\sum \gamma_h^d = 1$ that captures in some way what could be the hourly evolution of PV energy at day d. We will discuss in the next section how to define such a sequence using what we call an empirical clear sky PV energy curve. Assuming available such a table γ_h^d, we can disaggregate a daily energy prediction $\hat{E}^D$ into a 3-hour one at hour h as $\hat{E}_{d,h}^{D-3} = \hat{E}^D \Gamma_h^d$ with $\Gamma_h^d = \gamma_{h-2}^d + \gamma_{h-1}^d + \gamma_h^d$. We disaggregate this way the $\hat{E}_d^{D;y}$ and $\hat{E}_d^{D;5m}$ values into the corresponding 3-hour predictions $\hat{E}_{dh}^{D-3;y}$ and $\hat{E}_{dh}^{D-3;5m}$.

Table 1 summarizes the daily and 3-hour errors of all models considered in this section. As it can be seen, the 5-month models give the best results for both the daily and 3-hour predictions. Also, the 5-month daily and aggregated 3-hour model essentially tie for daily predictions. However, the 5-month 3-hour models give slightly better results for 3-hour predictions than the disaggregated daily model. Since we normalize productions to installed power, errors are given as percentages of installed power. Thus the average daily best error of about 28.4% corresponds to a total daily deviation of approximately 1.13GW and the 6.4% value of the 3-hour average error to about 256MW in three hours.

To end this section, we should mention that the SVR parameters C, ϵ and σ have been determined by five-fold monthly stratified CV. By this we mean that we build 5 different random folds for each month with basically six days at each. Instead of retaining just a single C, ϵ, σ set, we will keep the five optimal SVR parameter subsets that we use to build five different SVR models of each one of the types considered, i.e., yearly and 5-month models for daily and 3-hour predictions. Thus, the PV energy forecasts are actually the averages of the ones provided by these models.

4 Hourly PV Energy Forecasts

Recall that once we have the table γ_h^d, $1 \leq d \leq 365$, $0 \leq h \leq 23$, we can disaggregate the daily $\hat{E}_d^{D;y}$ and $\hat{E}_d^{D;5m}$ predictions into the $\hat{\mathcal{E}}_{dh}^{D-H;y}$ and $\hat{\mathcal{E}}_{dh}^{D-H;5m}$

Table 1. Day-ahead daily, 3-hourly and hourly errors of the yearly and 5-month models over the test months.

Daily model errors

Model	Dec13	Jan14	Feb14	Mar14	Apr14	May14	Ave
$e^{D;y}$	36.03	25.29	31.20	46.81	27.89	23.80	31.84
$e^{3H-D;y}$	44.71	28.36	34.33	48.30	29.80	28.43	35.66
$e^{D;5m}$	30.69	23.41	28.29	**37.12**	**29.22**	**21.84**	**28.43**
$e^{3H-D;5m}$	**22.58**	**23.19**	**27.06**	41.23	31.27	25.15	**28.41**

3-hour model errors

Model	Dec13	Jan14	Feb14	Mar14	Apr14	May14	Ave
$e^{3H;y}$	8.33	**6.17**	6.44	8.71	**6.12**	6.25	7.00
$e^{D-3H;y}$	6.50	6.57	8.18	10.59	7.50	6.51	7,64
$e^{3H;5m}$	**5.28**	6.53	**6.11**	**8.04**	6.68	**5.89**	**6.42**
$e^{D-3H;5m}$	5.77	6.53	7.87	9.52	7.72	6.31	7.29

Hourly model errors

Model	Dec13	Jan14	Feb14	Mar14	Apr14	May14	Ave
$e^{D-H;5m}$	2.28	2.72	3.15	3.79	3.06	2.52	2.92
$e^{3H-H;5m}$	**2.12**	**2.68**	**2.59**	**3.36**	**2.75**	**2.33**	**2.64**

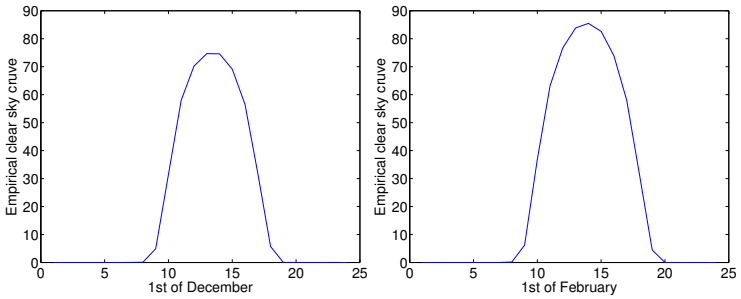
hourly ones simply as

$$\hat{\mathcal{E}}_{dh}^{D-H;y} = \gamma_h^d \hat{E}_d^{D;y},\ \hat{\mathcal{E}}_{dh}^{D-H;5m} = \gamma_h^d \hat{E}_d^{D;5m}.$$

Similarly, we can disaggregate the 3-hour predictions $\hat{E}_{dh}^{3;y}$ and $\hat{E}_{dh}^{3;5m}$ into the $\hat{\mathcal{E}}_{dh-j}^{3-H;y}, \hat{\mathcal{E}}_{dh-j}^{3-H;5m}$ hourly ones for the hours $h, h-1$ and $h-2$ as

$$\hat{\mathcal{E}}_{dh-j}^{3-H;y} = \frac{\gamma_{h-j}^d}{\sum_{k=0}^2 \gamma_{h-k}^d} \hat{E}_{dh}^{3H;y},\ \hat{\mathcal{E}}_{dh-j}^{3-H;5m} = \frac{\gamma_{h-j}^d}{\sum_{k=0}^2 \gamma_{h-k}^d} \hat{E}_{dh}^{3-H;5m},$$

for $j = 0, 1, 2$. Table 1 also contains the average errors of the 5-month hourly models considered. As it is to be expected, the best hourly results are those given by the best 3-hour model, i.e., the 5-month model. Now the average hourly best error of about 2.64% corresponds to an hourly error of approximately 105MW. Table 2 gives the average hourly errors for UTC hours 9 to 19. The largest error is 4.986% at hour 13, which corresponds to about 199MW. Figure 2 compares hourly PV production in the first week of December 2013, February 2014, April 2015 and the last one of May 2014 against the best hourly model and in Fig. 3 are depicted the MAE errors as a percentage of the installed power for the same period of time. As it can be seen, there is a relatively good agreement between



(a) December 1 (b) February 1

(c) April 1 (d) May 31

Fig. 1. Empirical clear sky PV curves for December 1, February 1, April 1 and May 31.

production and prediction values and sunrise and sunset are adequately handled.

We discuss next how we obtain the γ_h^d table, which we do using historical PV hourly energy values. The underlying idea is quite simple as we want the interpolating sequence γ_h^d to somehow reflect what would be the evolution of PV energy assuming day d to be a clear sky one. To do so, we first compute for each pair d, h a "maximum energy curve" μ_h^d defined as

$$\mu_h^d = \max_{y,q}\{\mathcal{E}_{d+q,h,y} : -\delta \leq q \leq \delta,\ y\}$$

where δ is some small integer and $\mathcal{E}_{d+q,h,y}$ denotes the energy produced at hour h in day $d+q$ of a year y. In other words, μ_h^d is the maximum of the normalized energy productions registered at hour h and any day in the interval $[d-\delta,\ d+\delta]$ over all years with PV energy production records (in our case from June 2011 to November 2013). We then smooth these μ_h^d values as

$$\rho_h^d = \frac{1}{2n+1}\sum_{-n}^{n}\mu_h^{d+m},$$

re-scale these ρ_h^d values to arrive to

$$\tilde{\gamma}_h^d = \frac{\max_j\{\mu_j^d\}}{\max_j\{\rho_j^d\}}\rho_h^d$$

and finally have

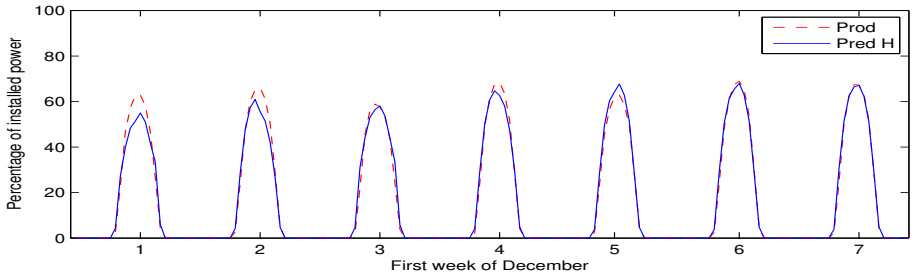
$$\gamma_h^d = \frac{\tilde{\gamma}_h^d / \sum_{h'}}{\tilde{\gamma}_{h'}^d}.$$

We have used the values $\delta = n = 10$ in our experiments. Figure 1 depicts the $\tilde{\gamma}$ curves for May 31, April 1, February 1 and December 1. When compared to actual maximum energy values, it is observed that the $\tilde{\gamma}_h^d$ curves clearly overshoot actual PV production, but since we interpolate using the normalized values γ_h^d, this effect is not important anymore.

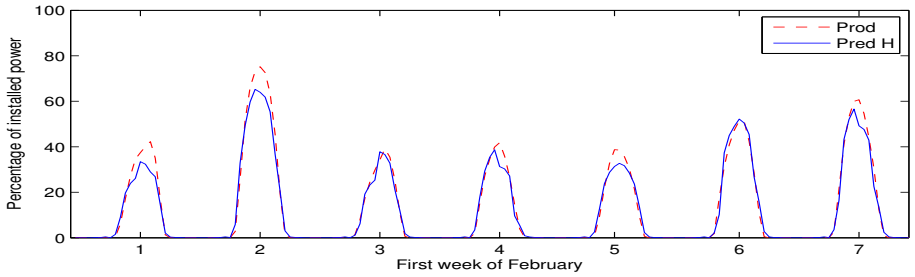
5 Same-day Day Hourly Updates of PV Energy Forecasts

Since NWP forecasts are updated usually only twice a day at around UTC hours 6 and 18, they cannot be used to update previous energy forecasts in an hourly basis. Other sources of information are needed and the most obvious one is the actual hourly production readings up to hour h that, in turn, can be used to obtain new forecast for hours $h+1, h+2, \ldots$ More precisely, for a given day d we may use energy readings $\mathcal{E}_{d6}, \ldots, \mathcal{E}_{dh}$ (we take PV production up to hour 5 as zero) to derive updated predictions $\hat{\mathcal{E}}_{dh+k}$ for the incoming hours $h+k, k = 1, \ldots$

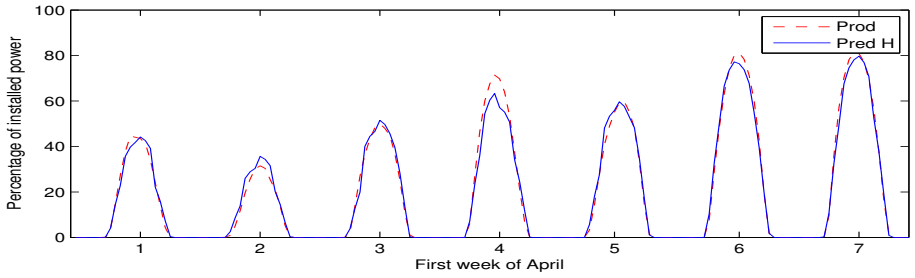
More information can be added, such as the errors e_{dh-j}, $j = 0, 1, \ldots$ of the day ahead predictions, the day ahead NWP forecasts or the day ahead prediction $\hat{\mathcal{E}}_{dh}^{da}$ in use for day d and hour h. However, we will just consider here the simplest approach of building at hour h a linear model $F_k^h(\mathcal{E}_{d6}, \ldots, \mathcal{E}_{dh})$ to approximate $\mathcal{E}_{dh+k}$. In other words, what we want is $\mathcal{E}_{dh+k} \simeq \hat{\mathcal{E}}_{dh+k} = F_k^h(\mathcal{E}_{d6}, \ldots, \mathcal{E}_{dh})$. Of course, the information in $v_h^d = (\mathcal{E}_{d6}, \ldots, \mathcal{E}_{dh})$ will be only relevant for h past enough from sunrise. Thus, the F_k^h models will be only relevant for $h_{SR}^d \leq h \leq h + k \leq h_{SS}^d$, with h_{SR}^d, h_{SS}^d the sunrise and sunset hours of day d. Moreover, the information in v_h^d will not be of interest for $h \simeq H_{SR}^d$. For these reasons and taking into account sun hours in Spain in the months between December and May, we take UTC hour 6 as sunrise and UTC hour 20 as sunset and thus we shall only consider models F_k^h built starting at hour $h = 9$ and ending their predictions at hour $h + k = 19$.



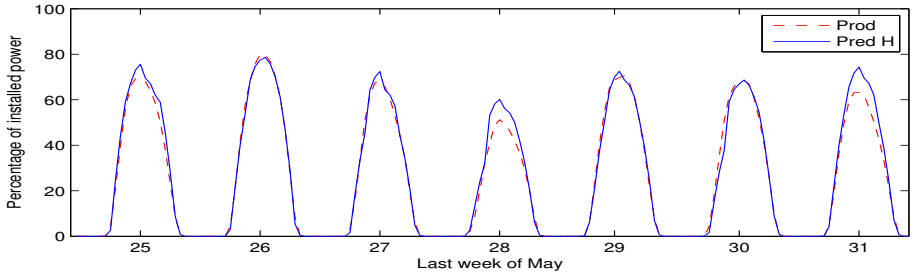
(a) December 1–7 2013



(b) February 1–7 2014

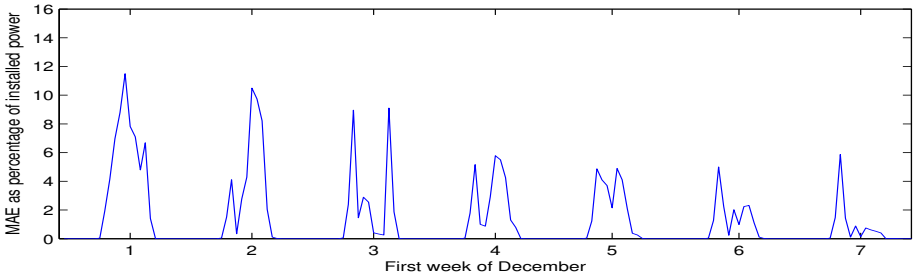


(c) April 1–7 2014

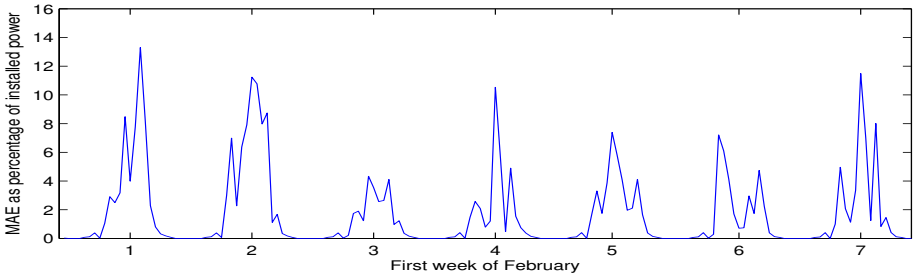


(d) May 25–31 2014

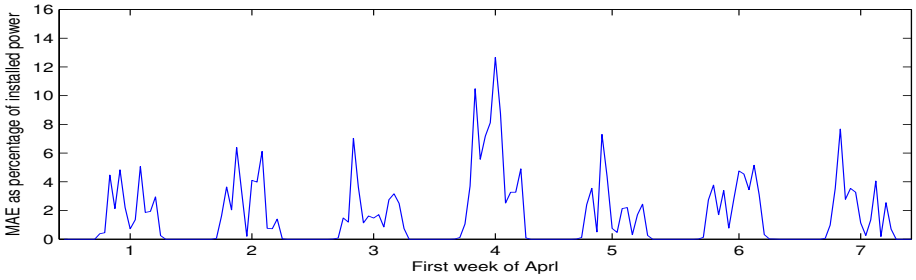
Fig. 2. Hourly prediction (red) vs production (blue) % of installed PV capacity on the weeks starting at December 1, 2013, February 1, April 1, and May 25, 2014.



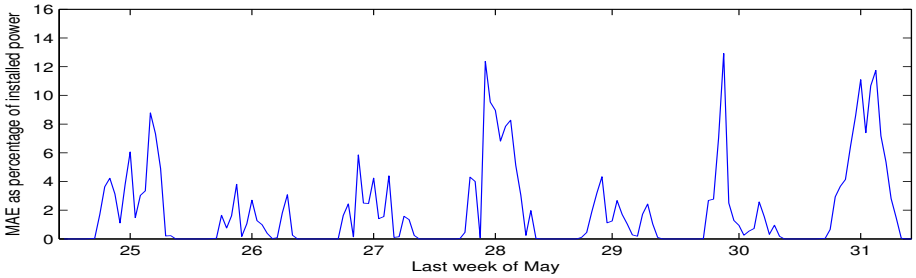
(a) December 1–7 2013



(b) February 1–7 2014



(c) April 1–7 2014



(d) May 25–31 2014

Fig. 3. MAE errors as a % of installed PV capacity on the weeks starting at December 1, 2013, February 1, April 1, and May 25, 2014.

As mentioned in Sect. 2, we will use RR to build the F_k^h models. Since these models do not need NWP information, we can use the entire period from June 2011 to November 2013 for training and validation purposes. As done before, we have built these models using for training past 5-month periods centered in the month that we use for testing. We estimate the λ_k^h parameter of each RR model using for validation 5-month periods between December 2012 and November 2013. Notice that the maximum number of model parameters is about $18 - 6 + 1 = 13$ while the number of samples is much higher. Thus the λ_k^h have very small values and, in fact, models built using standard linear regression or Lasso give very similar results.

We report our results in Table 2. The top row gives the MAE errors of the best day ahead hourly model (that we recall was given by the 3-hour 5-month SVR models) and the other rows give the average errors over the testing period of the F_k^h models with $10 \leq h \leq 19$. We mark in bold face the hours where the F_k^h models beat the day-ahead one. As it can be seen this is the case on a band that contains about 3–4 subdiagonals. Figure 4 captures this behavior, and in Fig. 5

Table 2. MAE errors of intra-day hourly updates as % of installed PV capacity versus the best DA model. Columns represent predicted hours, and rows the hour in which a new prediction is issued.

		H9	*H10*	*H11*	*H12*	*H13*	*H14*	*H15*	*H16*	*H17*	*H18*	*H19*
	DA	4,533	3,552	4,023	4,575	4,986	4,558	3,963	4,204	2,696	1,467	0,384
	H6	9,638	11,570	12,843	13,368	13,573	13,391	12,507	10,326	7,500	3,046	**0,361**
	H7	7,086	9,586	10,820	11,492	11,653	11,509	10,810	8,790	6,358	2,233	**0,372**
	H8	**3,348**	6,772	8,693	9,709	10,127	10,195	9,444	7,485	5,079	1,692	**0,339**
	H9	-	**1,865**	**3,411**	**4,482**	5,338	6,073	6,101	5,785	4,666	1,757	**0,337**
HU	*H10*	-	-	**1,014**	**2,171**	**3,313**	**4,353**	4,992	5,269	4,727	1,691	**0,341**
	H11	-	-	-	**1,044**	**2,254**	**3,322**	4,062	4,655	4,136	1,576	**0,351**
	H12	-	-	-	-	**1,139**	**2,330**	**3,382**	4,305	4,155	1,556	0,403
	H13	-	-	-	-	-	**1,142**	**2,274**	**3,584**	3,707	1,525	**0,334**
	H14	-	-	-	-	-	-	**1,151**	**2,637**	3,361	**1,393**	**0,336**
	H15	-	-	-	-	-	-	-	**1,303**	**2,464**	**1,290**	**0,373**
	H16	-	-	-	-	-	-	-	-	**0,997**	**0,946**	**0,319**
	H17	-	-	-	-	-	-	-	-	-	**0,540**	**0,307**
	H18	-	-	-	-	-	-	-	-	-	-	**0,226**

it can be also observed the performance of this model for some specific days. More concretely, in this last image it is compared the real production against the result of the models obtained from hour 9h and from hour 12h.

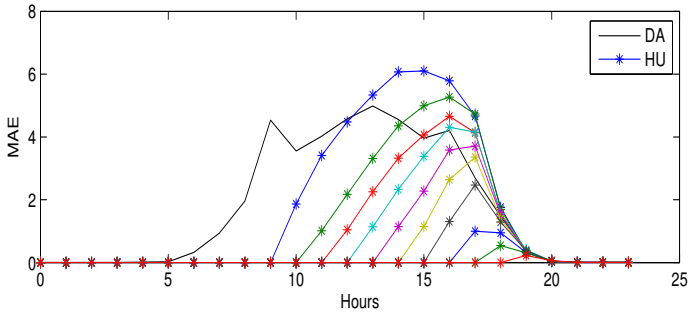
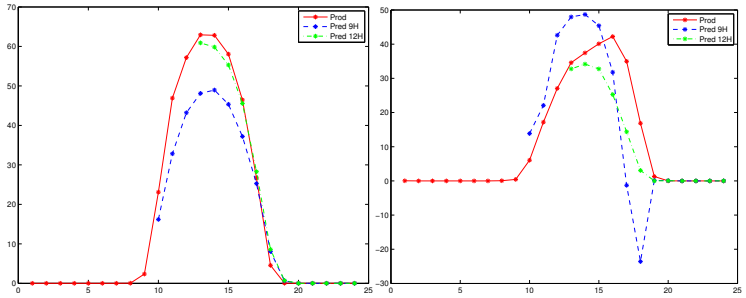
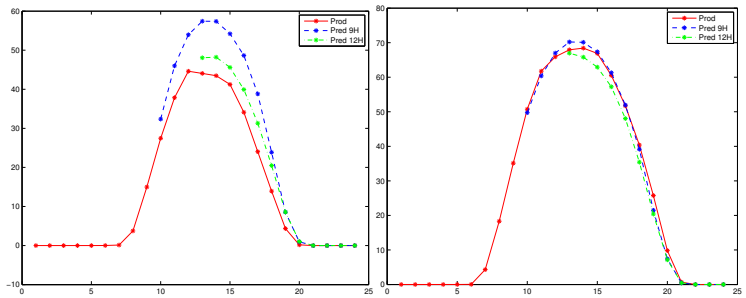


Fig. 4. MAE errors as % of installed PV capacity of same-day hourly update (HU) versus those of the day-ahead hourly models (DA). Different colors identify the errors of the HU models built at consecutive hours after the first one at $h = 9$.



(a) December 1 (b) February 1



(c) April 1 (d) May 31

Fig. 5. Behavior of HU models obtained from hour 9 and from hour 12 for December 1, February 1, April 1 and May 31.

6 Discussion and Conclusions

Wide area prediction of PV energy is an important issue for the Transmission System Operators (TSO) of countries such as Spain where there is a very large number of small and geographically dispersed PV installations. Small individual outputs and dispersion mean that the impact of a single installation is not all that relevant. However, the aggregated effect of all the PV installations is quite important (about 4 GW of peak power in Spain and much higher in other European countries) and an accurate prediction of the global PV output is of a very high interest.

In this work we have addressed day-ahead and intra-day PV energy predictions for peninsular Spain. The inputs in the first, day-ahead case are NWP forecasts from the ECMWF and we derive at day d PV predictions for the entire day $d+1$ at basically two extremes: daily aggregated predictions and individual hourly values. In the second, same-day case, the inputs are the production at a given day up to UTC hour h, $9 \leq h \leq 18$, from which we derive forecasts for UTC hours $h+1, \ldots, 19$. From a ML point of view, both problems are fairly simple as we only have to map input data into accurate predictions. SVR was used for day-ahead models, either as a single model built over an entire year or as month-tailored models built over 5 months, with the target one as the center. These 5-month models give the best day ahead results for the prediction of daily aggregated PV energy and its 3-hour values. Since 3-hour is the smallest resolution available using NWP, to arrive at hourly predictions we disaggregated 3-hour forecasts using what we called empirical clear sky PV energy curves.

To derive intra-day hourly forecasts we have used plain Ridge Regression (RR) models. Applying them from hour 9 onwards, these models' predictions beat the day ahead ones in the next 3–4 hours. Given the simplicity of the approach followed, this is an encouraging result that, however, should be improvable possibly adding more information to the models to be built.

In summary, the paper shows how relatively simple ML techniques coupled with modeling considerations derived from the periodical nature of radiation, and the possibly different response of PV cells at different hours can be applied to obtain good forecasts of the PV energy produced over a large area. Of course, while accurate individual PV plant forecasting is much harder (in part because of the current NWP resolution and their physical modeling of radiation-related phenomena), large scale PV modeling takes advantage of the powerful smoothing derived from large area energy aggregation. However, this is still an important problem for electrical TSOs where further work is clearly needed. Besides extending the six month results here over an entire year, possible issues are a better modeling of the empirical clear sky PV energy curves, a better understanding of the nature of the modeling errors (SVR modeling tends to undershoot mid-day energy production) and of the effect on them of cloudy days, or the improvement of the intra-day hourly forecasts. On the other hand, other ML approaches could be used. For instance, in day-ahead predictions we could go from the simplest case of regularized linear models to the more complex one of using, say, Multilayer Perceptrons, possibly after a dimensionality reduction process that

provides more manageable input patterns. Moreover, the estimation of forecast uncertainty is also important for TSOs. Theoretical uncertainty estimates are well known for linear models and also do exist for SVR [8,3]. We are working on these and other related issues.

Acknowledgments. With partial support from Spain's grants TIN2010-21575-C02-01 and TIN2013-42351-P, and the UAM–ADIC Chair for Machine Learning. We thank Red Eléctrica de España for useful discussions and making available PV energy data.

References

1. Benghanem, M., Mellit, A., Alamri, S.: Ann-based modelling and estimation of daily global solar radiation data: A case study. Energy Conversion and Management 50(7), 1644–1655 (2009)
2. Chang, C., Lin, C.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
3. Chu, W., Keerthi, S.S., Ong, C.J.: Bayesian support vector regression using a unified loss function. Neural Networks, IEEE Transactions on 15(1), 29–44 (2004)
4. ECMWF: European Center for Medium–range Weather Forecasts. `http://www.ecmwf.int/` (2005)
5. GFS: NOAA Global Forecast System. `http://www.emc.ncep.noaa.gov/index.php?branch=GFS` (2014)
6. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Verlag (2001)
7. Kleissl, J.: Solar Energy Forecasting and Resource Assessment. Academic Press (2013)
8. Lin, C.J., Weng, R.C.: Simple probabilistic predictions for support vector regression. Techical Report, Department of Computer Science, National Taiwan University (2003)
9. Mellit, A., Kalogirou, S.A.: Artificial intelligence techniques for photovoltaic applications: A review. Progress in Energy and Combustion Science 34(5), 574–632 (2008)
10. Myers, D.R.: Solar radiation modeling and measurements for renewable energy applications: data and model quality. Energy 30(9), 1517–1531 (2005)
11. Pedro, H., Coimbra, C.: Assessment of forecasting techniques for solar power output with no exogenous inputs. Solar Energy 86, 2017–2028 (2012)
12. Pedro, H., Coimbra, C.: Stochastic learning methods. In: Kleissl, J. (ed.) Solar Energy Forecasting and Resource Assessment. pp. 383–407. Academic Press (2013)
13. Scholkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press (2001)
14. Ulbricht, R., Fischer, U., Lehner, W., Donker, H.: First steps towards a systematical optimized strategy for solar energy supply forecasting. In: Proceedings of the DARE 2013, Data Analytics for Renewable Energy Integration Workshop. pp. 14–25 (2013)
15. Wolff, B., Lorenz, E., Kramer, O.: Statistical learning for short-term photovoltaic power predictions. In: Proceedings of the DARE 2013, Data Analytics for Renewable Energy Integration Workshop. pp. 2–13 (2013)