Short-Term Spatio-Temporal Forecasting of Photovoltaic Power Production

Xwégnon Ghislain Agoua, Robin Girard, Georges Kariniotakis

Short-Term Spatio-Temporal Forecasting of Photovoltaic Power Production

Xwégnon Ghislain Agoua, Robin Girard and George Kariniotakis, *Senior Member, IEEE*

***Abstract*—In recent years, the penetration of photovoltaic (PV) generation in the energy mix of several countries has significantly increased thanks to policies favoring development of renewables and also to the significant cost reduction of this specific technology. The PV power production process is characterized by significant variability, as it depends on meteorological conditions, which brings new challenges to power system operators. To address these challenges it is important to be able to observe and anticipate production levels. Accurate forecasting of the power output of PV plants is recognized today as a prerequisite for large-scale PV penetration on the grid. In this paper, we propose a statistical method to address the problem of stationarity of PV production data, and develop a model to forecast PV plant power output in the very short term (0-6 hours). The proposed model uses distributed power plants as sensors and exploits their spatio-temporal dependencies to improve forecasts. The computational requirements of the method are low, making it appropriate for large-scale application and easy to use when on-line updating of the production data is possible. The improvement of the normalized root mean square error (nRMSE) can reach 20% or more in comparison with state-of-the-art forecasting techniques.**

***Index Terms*—Autoregressive processes, forecasting, photovoltaic systems, smart grids, spatial correlation, stationarity, time series.**

I. Introduction

GROWING global energy demand and increased awareness of the consequences of climate change have put renewable energy in the spotlight. Renewable energy generation, and particularly photovoltaic (PV) energy, is continuously increasing in several countries, especially in Europe. The power output of a PV plant depends on meteorological conditions. In regions subject to active weather changes, it is characterized by high variability and low short-term predictability. These characteristics challenge power system operators, since they introduce uncertainties into the various functions of power system management, especially for large-scale PV integration.

The PV production expected in the next few minutes, hours or days needs to be accurately forecasted in order to efficiently perform functions like scheduling power systems, minimizing reserve costs [1], trading PV production in electricity markets and coordinating PV plants with storage, and in general to contribute to increasing the competitiveness of renewable energy technologies [2]. In the context of smart grids, PV forecasts are necessary to manage distribution networks, microgrids or smart homes, where other options like active demand, storage, electric vehicles etc., coexist with PV generation [1], [3].

The literature proposes several methods to forecast PV production [4]. These methods can be classified according to their specific forecast horizon [5]. The final choice of forecasting technique is related to this horizon and the available data. The most common statistical methods are regression methods like linear regression, regression trees, boosting, bagging, random forests, Support Vector Machines [6]–[9], and semi-parametric models. These techniques investigate the correlation between the historical production and the related meteorological measurements [10]. The Box and Jenkins time series treatment methods (ARIMA, ARMA, SARIMA, ...) are also widely used in PV power forecasting. The question of the series stationarity is treated by pre-processing steps using either clear sky modeling, [11]–[13] or certain normalization techniques employing Top of Atmosphere (TOA) or Global Horizontal Irradiance (GHI). In [14], [15], regression-based methods are also used. Data mining techniques are employed to cluster past events into historical data on production and/or meteorological variables. This same idea of similarity is used to forecast production when PV panels are covered by snow [16].

Neural networks have been used to forecast PV production with different types of activation functions [17]. They are often compared or coupled with physical models [18], [19]. They can also be used as a second step in a two-step modeling chain, where the first step is to predict meteorological variables using Numerical Weather Predictions (NWP) [20], [21].

Recent years have seen increasing interest in techniques that can take into account not only historical data about the site that is the object of forecasting, but also other spatially distributed data. These methods, initially proposed for wind power forecasting, are developed for different applications, like identifying regions with high energy production potential [22], [23], studying the spatial propagation of forecasting errors [24], [25], and even "geographically intelligent" prediction [26]–[28].

Most references refer to spatio-temporal solar irradiation forecasts. Spatial information from sky cameras or satellite images is used and described in 2D or even 3D with cloud motion vectors. Cloud movement predictions lead to solar radiation forecasts for very short-term horizons (a few minutes up to 2-4 hours ahead) [29]–[31]. NWP models and cloud motion vectors can also be combined for short-term forecasts (a few hours up to 2 or 3 days ahead) [32]. Solar radiation can also be forecasted with auto-regressive models in time and

The authors are with MINES ParisTech, PSL Research University, PERSEE - Centre for Processes, Renewable Energies and Energy Systems CS 10207 rue Claude Daunesse, 06904 Sophia Antipolis Cedex, France. (e-mails: xwegnon.agoua@mines-paristech.fr, robin.girard@mines-paristech.fr, georges.kariniotakis@mines-paristech.fr)

Manuscript received February 13, 2017; revised May 31, 2017 and July 20, 2017; accepted August 26, 2017.

space or kriging [33]–[36]. These methods employed in solar radiation forecasts can be costly due to the complexity of the required measuring infrastructure and data, and the modeling chain that has to be developed.

In this paper we propose a forecasting methodology that exploits the spatial and temporal correlations in existing data from geographically dispersed PV installations to predict the power output of a specific plant. Short-term forecast horizons of a few minutes up to 6 hours are considered. The models investigated here directly use geographically dispersed power plants as a network of sensors. This differentiates the approach from methods that use off-site data from meteorological stations and ground-based irradiance sensors as in [37]. The proposed model does not consider input from a NWP model, and forecasts are made based on the production data and not global irradiance data as is the case in [38], [39].

In a preceding conference paper [40], the authors have proposed a spatio-temporal methodology. In this paper that methodology is significantly improved on several points. The first improvement consists in proposing a new stationarization process, that unlike [41], does not involve modeling for the clear sky generation. The proposed approach aims to overcome weaknesses of the clear sky based normalization especially for early and late hours of the day when solar irradiation is low. The second improvement proposed here permits to take into account the local meteorological conditions in the spatio-temporal model. This is done by defining model coefficients dependent on the weather variables in the estimation process. The third improvement is to propose a model that integrates an automatic selection of the appropriate input variables. This is particularly adapted to highly dimensional problems, as can be the case for spatio-temporal PV forecasting. Finally, the spatial density of the considered PV plants in real-world cases can be variable, and for this reason we illustrate the usefulness of the proposed methodology with two test cases featuring a low and high number of PV plants. The dimensionality problem and the importance of the proposed variable selection process are highlighted through the test case with high number of PV installations (185 PV plants). The benefits in terms of performance of all the above contributions with respect to [40] are presented in section IV-C.

The paper is structured as follows: the potential of making use of spatio-temporal information is investigated in section II with a focus on the proposed stationarization procedure, the data and the evaluation criteria. The proposed spatio-temporal models are presented in section III. The results are presented and discussed in section IV. Finally, the conclusions of the study are discussed in section V.

II. Analysis of the Interest of Spatio-temporal Modeling

The aim in this section is to demonstrate the interest of using spatio-temporal information for PV forecasting purposes. This is done through an analysis of the correlations between data from PV plants. However, given that these data are dominated by the daily sun cycle, which biases correlation analysis, it is necessary to subtract the periodic components in the series through appropriate stationarization. To achieve this, we propose a new method to stationarize PV production series that is not only useful for analyzing correlations but also for building the forecasting models themselves. Initially, two test cases are introduced that provide real world data, used in this section to assess the proposed stationarization process, and in later sections to evaluate the proposed forecasting models.

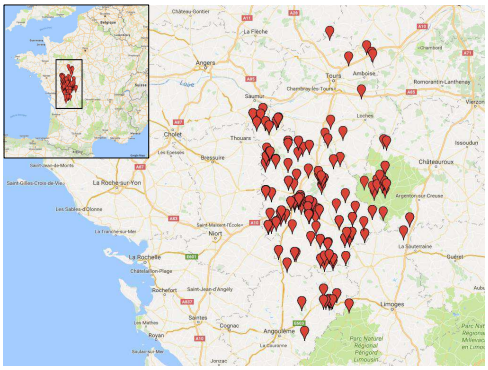

Fig. 1. The power plants of the second data set d_2 located in west central France. The distance between the power plants ranges from 1 km to 230 km.

A. Test Cases

Two data sets are considered in this paper corresponding to relatively different climatic conditions and different spatial densities of installed PV plants as well as the distances between them. The first data set labeled d_1, consists of time series of the measured PV generation of a set of 9 power plants located in the south of France. The power plants are labeled P1-P9 with peak power ranging from 45 kWc to 5 MWc. The distance between the power plants ranges from 5 km to 465 km. The measurements cover a period of 20 months starting from July 2013 with a resolution of 6 min to 15 min depending on the PV plant. The data quality has been checked to remove inconsistencies and then interpolated to produce series with a 15 min temporal resolution that are used hereafter.

The second data set labeled d_2 is a good illustration of a case featuring a high number of power plants and significant geographic density. It comprises the output of 905 PV power inverters in the mid-west region of France with peak power ranging from 3.2 kWc to 58 kWc. This amount of inverters corresponds to 185 different PV power plants (set of inverters at the same location). The distance between them varies from 1 km to 230 km. The data relate to the period from November 2014 to March 2016. The original time resolution of the data is 5 min, which was averaged to produce series with a 15 min temporal resolution as with the previous test case. The locations of the power plants in the test case d_2 are represented in Figure 1.

B. The Stationarization Procedure

Most of the time series analysis methods require stationary series. The photovoltaic production series are not stationary because the average production depends on the time of day, while the variability, as expressed by the variance of the

production, depends on the level of production and indirectly on the time of day. A simple differentiation of the series is not efficient in producing stationary time series because the non-stationarity in the variance remains.

Here we propose a procedure to stationarize a PV production series. The aim is to decompose the production series using a deterministic component that describes the movement of the sun. It is inspired from the clear sky index for solar radiation [42]–[44].

The clear sky index for solar radiation represents the way that the atmosphere attenuates light on an hour-to-hour or day-to-day basis as a function of the movement of the earth around sun. It is defined as the quotient of radiation actually measured by the radiation simulated with a clear sky model.

This index makes it possible to remove the diurnal and seasonal pattern from irradiation data, which is expected to improve the performance of the statistical techniques applied thereafter. Here, we define it as the ratio between irradiation measurements and an advanced clear sky estimate at time t:

$$k_t^{irr} = \frac{I_t^{meas}}{I_t^{sim}}.$$

In a similar way, we define a clear sky index for photovoltaic power k^{pv} as

$$k_t^{pv} = \frac{P_t^{meas}}{P_t^{sim}} \tag{1}$$

where P_t^{meas} is the PV production measured at time t, P_t^{sim} is the simulated production output for time t. P_t^{sim} is constructed as the product of the PV overall system efficiency parameter η and the simulated irradiation I_t^{sim} at either the Top of Atmosphere (ToA) level or under clear sky conditions as proposed by the European Solar Radiation Atlas (ESRA) model [43]. The parameter η embedded the efficiency of the generator and the active surface.

Although intuitively, the index k_t^{pv} would be expected to be adequate for de-trending, in practice appropriate stationarity tests on the resulting series (i.e. unit roots tests) indicate that the results are not satisfactory. For this reason we propose a new relation between the actual production and P_t^{sim} using a function f that would explain more accurately the link between the two productions. This function would also help to reduce the non-stationarity when defining the new working series u_t for the hours at which P_t^{sim} is not zero as

$$u_t = P_t / f(P_t^{sim}). \tag{2}$$

The irradiation considered for defining P_t^{sim} is the simulated ESRA series as it embeds more information about the atmospheric characteristics than the ToA, such as albedo, air mass, the Linke turbidity factor and other atmospheric conditions. The simulation of irradiation was done under the hypothesis of a horizontal surface; this is because the inclination does not affect the stationarization since the variation in the output level it produced would be assimilated by η. Different types of relation can be conceived for f including linear, quadratic and piecewise linear.

The choice of the appropriate function was made using a quantitative criterion based on the evolution of the daily standard deviation of the series u_t. The retained function is piecewise linear in the simulated production and depends on the direction of the productions daily evolution (either increasing at the beginning of the day or decreasing after solar noon). It can be expressed as:

$$f(P_t^{sim}) = P_t^{sim} + f_a(P_t^{sim}) + f_b(P_t^{sim}) \tag{3}$$

where the function f_a is defined from sunrise to noon and f_b from noon to sunset. The goal of f_a and f_b is to improve the treatment at the beginning and at the end of the day. Their definition on a daily basis is:

$$\left\{\begin{array}{rcl} f_a(0) & = & \alpha_a \\ f_a\left(\beta_a \frac{P_{max}^{sim}}{2}\right) & = & 0 \\ f_a(P_{max}^{sim}) & = & \gamma \end{array}\right. \quad \left\{\begin{array}{rcl} f_b\left(P_{max}^{sim}\right) & = & \gamma \\ f_b\left(\beta_b \frac{P_{max}^{sim}}{2}\right) & = & 0 \\ f_b(0) & = & \alpha_b \end{array}\right. \tag{4}$$

where P_{max}^{sim} is the maximum production simulated for the day. The values of the coefficients $\alpha_{a,b}, \beta_{a,b}, \gamma$ are obtained through an optimization process that aims to minimize the standard deviation criterion. The optimization is made under the constraints $\beta_{a,b} \in (0, 2)$. The coefficients are randomly initialized and then optimized considering a sliding window of one-month over the ESRA irradiation time series. The sliding window covers the period prior to the day of interest. The stationarity of the normalized form of u_t was evaluated by analyzing its autocorrelogram and computing unit root test.

The procedure can be summarized in the following steps for a power plant:

1) Clean the spurious data from the PV production series.
2) Simulate the ESRA clear sky irradiation series and the corresponding power series using the plant's efficiency.
3) Determine the appropriate coefficients of the functions (f_a, f_b) using an optimization process on a sliding interval of simulated irradiation values.
4) Normalize the measured series P_t^{meas} to obtain the series u_t.

C. Analysis of Spatial Correlations

To investigate the existence of spatio-temporal patterns, we evaluate the cross-correlation between the lagged production series. However, this requires eliminating the effect of East to West correlation transfer by considering the stationarized series for the PV plants.

Figure 2 presents the empirical cumulative distribution of the cross-correlation values for the power plants in the data set d_2. Three distributions are plotted for three classes of distance between the power plants (from the closest to the farthest). The figure shows that the cross-correlation values are higher for the first class of distance (less 50 km) than for the last class (more than 100 km). As the effect of the bell-shape in the stationarized production data is absent, we can assume that the link described by these correlation values is due to a spatial transfer of information between the power plants mainly due to cloud movements. This analysis confirms the interest of a forecasting solution that takes into account both the temporal and spatial variability of the production series.

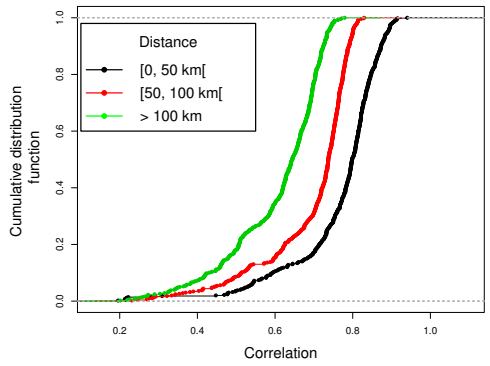

Fig. 2. Data set d_2: Cumulative distribution function (CDF) of the cross-correlation between the lagged production series. Green, red and black CDFs respectively described three classes of distance between the power plants.

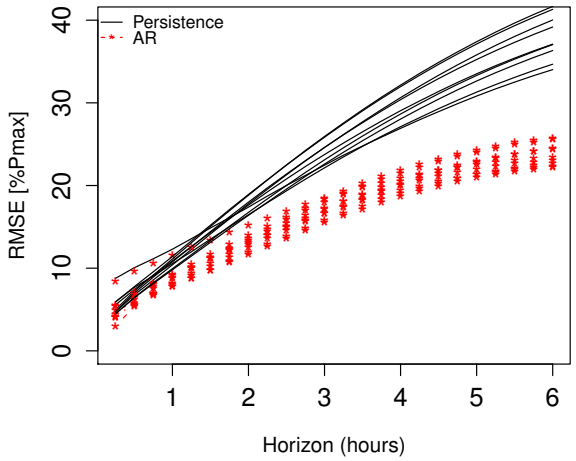

Fig. 3. Data set d_1: Comparison of the normalized RMSE of AR and persistence models over the testing set. Solid and dotted lines represent respectively the performance of persistence AR models. The forecast time step is 15 min.

III. A Model for Spatio-Temporal PV Forecasting

A. The Reference Model

In order to be able to compare the advantages of a spatio-temporal approach for PV forecasting, we introduce reference models for benchmarking that does not use such geographically distributed information. Several methods can be used to forecast PV generation as presented in the introduction. The persistence model is often used as a reference in the literature on renewable energy forecasting to compare the performance of advanced models, as it is easy to compute, is based only on measured data, and does not involve any modeling processes. Thus, the persistence results are easily replicable. Moreover, in practical applications of PV forecasting, persistence is often chosen as a fallback model to provide forecasts in case advanced models fail. We define here as persistence a model that considers that the power production of a PV plant at time $t+h$ is the same as the production of that plant at the same time on the previous day. This approach does not consider any off-site data. Despite its popularity as a reference model in the literature, its overall performance is poor [4]. To account for the different factors that affect PV production one could adjust persistence as a function of the observed values on the current day. However, this already involves some data manipulation, and different options could be considered, but such empirical adjustments are out of the scope of this paper. To avoid obtaining overoptimistic results from a spatio-temporal method, it is also necessary to use an advanced reference" model featuring state-of-the-art performance and reasonable complexity so that results can be easily reproduced. For this purpose we consider autoregressive (AR) models described as:

$$\hat{P}^x_{t+h|t} = \hat{\beta}^0_h + \sum_{l=0}^{L} \hat{\beta}^l_h P^x_{t-l} \tag{5}$$

where P^x_t is the production of the power plant x at time t and $\hat{P}^x_{t+h|t}$ the prediction for horizon h. The appropriate maximum time lag L is chosen by minimizing the Akaike Information Criterion (AIC). We applied this model to the data set d_1 using 15 months for learning and 5 months for the tests. Forecasts are updated at each 15-minute time step.

For the 5 months of testing set for each power plant, we also applied the persistence and compared its performance with the AR model. Figure 3 presents the normalized root mean square error RMSE for the AR and persistence models for the d_1 power plants as a function of the prediction horizon. The figure shows that the best model is the AR model, as its RMSE levels are the lowest. We thus retain the AR model as a reference in this paper to evaluate the performance of the spatio-temporal forecasting models. With our reference model thus defined, we can evaluate the contribution of integrating additional information from neighboring plants.

B. The Proposed Spatio-Temporal Model

The correlation analysis carried out in subsection II-C confirms the interest of using measurements from other power plants to increase the quality of the PV power forecasts. We propose here a spatio-temporal model that produces PV power forecasts for a power plant using measurements from other plants nearby.

Let $\mathcal{X}$ be the set of N PV plants and Ls the appropriate maximum lag. The forecast model for a power plant of interest x is then defined as:

$$P^x_t = \beta^0 + \sum_{l=0}^{Ls} \sum_{y \in \mathcal{X}} \beta^{l,y} P^y_{t-l} \quad . \tag{6}$$

For a selected horizon h, the coefficients $\beta = (\beta^0, \beta_r)$ with $\beta_r = (\beta^{l,y})_{0 \le l \le Ls,\, y \in \mathcal{X}}$ are estimated using a least squares method that involves minimizing the Residual Sum of Squares (RSS):

$$RSS(\beta) = \|\mathbf{P^x} - \mathbf{X}\,\beta\|^2, \tag{7}$$

where $\mathbf{P^x}$ is the measurement for power plant x.

$\mathbf{X}$ is a $N \times (Ls+1)$ matrix the lines of which are the current and lagged production for the power plants y_i

$$\mathbf{X} = \begin{pmatrix} 1 & P_t^{y_1} & \dots & P_{t-Ls}^{y_1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & P_t^{y_N} & \dots & P_{t-Ls}^{y_N} \end{pmatrix}. \quad (8)$$

The forecast at time t for the horizon h for a power plant x is then defined by:

$$\hat{P}_{t+h|t}^{x} = \hat{\beta}_h^0 + \sum_{l=0}^{Ls} \sum_{y \in \mathcal{X}} \hat{\beta}_h^{l,y} P_{t-l}^{y}. \quad (9)$$

The first issue related to the above model is the dimensionality problem when there is a high number of PV plants. To reduce the complexity of the model in such cases, we propose a two-step variables selection procedure. Let us call x the power plant of interest for which the forecasts are made. The first step is to compute the distance between the plant x and the other plants and select the n_p closest plants to x. The second step is to apply a stepwise selection procedure based on the AIC criterion. The selection is made backward; the n_p variables and their respective lags are integrated into the model and then removed one by one and the AIC is recalculated each time. The model with the minimum AIC is retained.

C. Extension of the Model: Spatio-Temporal Model using Clusters of Meteorological Conditions

The previous model is purely based on the historical production data. Here, we propose a variant of the model that allows a smooth dependency of the linear model coefficients on local meteorological conditions. The meteorological variables can be temperature, wind speed or direction, or another variable. These measurements are obtained from the closest weather station. With the previous notation, the forecast for the horizon h is denoted as:

$$\hat{P}_{t+h|t}^{x} = \hat{\beta}_h^0(Z) + \sum_{l=0}^{L} \sum_{y \in \mathcal{X}} \hat{\beta}_h^{l,y}(Z) P_{t-l}^{y}, \quad (10)$$

where Z represent the meteorological variables. The coefficients are estimated by weighted least square regression by:

$$\hat{\beta}_h^{l,y}(z) = Arg\min \sum_t \phi\left(\frac{Z_{t,y} - z}{\gamma}\right)\left(P_t^y - P_{t+h}^y\right)^2 \quad (11)$$

where the coefficient γ is the mean of the random variable Z. The weights function is exponential:

$$\phi(x) = \exp\left(-||x||^2/2\right). \quad (12)$$

The weights are calculated using the measurements from the closest available meteorological station.

D. Improved Variable Selection Procedure

In the model presented above, the dimensionality problem (i.e. high number of variables) is treated with a simple selection variable procedure. The model can be modified to directly treat the variable selection issue using LASSO. The Least Absolute Shrinkage and Selection Operator [45] regression integrates a penalty into the minimization problem by applying a constraint on the sum of the absolute values of the coefficients. The estimator is defined as:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} RSS(\beta) + \lambda \|\beta\|_1 \right\}. \quad (13)$$

Some bias is introduced but the variance is reduced. The selection of the coefficients is automatic and some of them are set to zero for high values of the penalization parameter λ. The regularization parameter λ is obtained by cross-validation and the path of the solutions of β is piecewise linear in λ.

IV. Evaluation

The proposed models are applied to the data sets d_1 and d_2 for a 6-hour horizon with a 15-min time step, and with a sliding window scheme that updates forecasts every 15 min. The forecasts are compared to those of the reference model. The models were developed using the software R [46].

A. Impact of the Stationarity Procedure on Forecast Errors

The reference AR model was applied to the two types of production series of d_1: the raw series and the series that was stationarized following the procedure proposed in Section II. The RMSE for the respective series was computed for each power plant and the improvement due to the stationarization was calculated. For all the plants except P4, there is a significant improvement in RMSE when stationarized series are used. The average improvement in terms of RMSE is 7%. The stationarized series perform better than the raw inputs. The case of P4 can be explained by the fact that the AR model efficiently captures the temporal variability with standard normalization.

The same analysis was made of the power plants in data set d_2, where 136 power plants were retained after data cleaning. Figure 4 represents the improvement of the RMSE achieved with the stationary procedure for d_2. The mean improvement for 3-hour horizons is 10% and can reach 15%. This significant reduction in forecasting errors confirms the efficiency of the stationarization method and the interest of using it to pre-process data before integrating them into the forecasting model. Thus hereafter, we use the stationarized series.

B. Performance of the Spatio-Temporal Model

For each of the power plants of d_1, we apply the spatio-temporal model in its form defined in part III-B. The standardized errors are computed at time t for look-ahead time h ranging from 15 min to 6 hours. The densities of the prediction errors are computed using kernel density estimation and are presented in figure 5 for two power plants and different

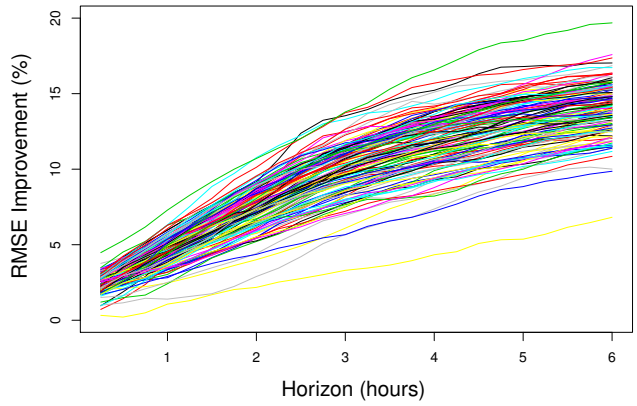

Fig. 4. Data set d_2: RMSE Improvement of the AR model with stationary series over non-transformed data. Each line represents the improvement obtained for a power plant. The time step is 15 min.

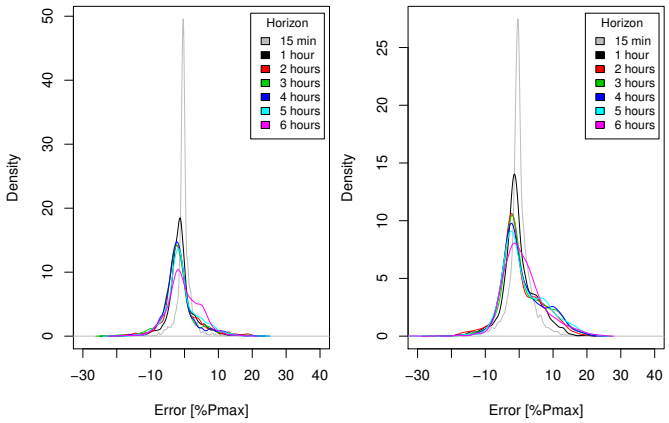

Fig. 5. Densities of the forecasting errors post spatio-temporal model for 2 power plants (Kernel estimation). The horizons range from 15 min to 6 hours.

horizons. Note that for both power plants the distributions are not Gaussian, as the modes and averages are significantly different. The averages are close to zero and the skew is negative. As the horizon increases, the distribution mode shifts to the left. The same analysis was performed on the other power plants of d_1 with the same conclusions.

To obtain a more complete overview of the proposed models performance, we compare it to random forest (RF) models. RF models are shown in the literature [10] to be one of the most efficient models to produce accurate forecasts of PV power production. We thus computed an RF model and compared its performance to the spatio-temporal model. Table I presents the minimum, mean and maximum RMSE improvement over the 6-hour time horizons of the spatio-temporal model (ST) w.r.t. the AR and RF models for a sample of five power plants of d_1. The table shows an average improvement of around 10% for the ST model compared with the AR model and 6% compared with the RF one. The improvement compared to the AR and RF models can reach respectively 20% and 15%. The improvement values are quite similar for all of the power plants except for plant P8, for which there is no improvement. This is the most distant power plant, and the spatial correlation does not reach it.

TABLE I
RMSE IMPROVEMENT OF THE SPATIO-TEMPORAL (ST) MODEL OVER THE REFERENCE AR MODEL AND THE RANDOM FOREST (RF) MODEL FOR 5 POWER PLANTS OF DATA SET d_1.

Improvement of RMSE (%)		P1	P2	P4	P5	P6
ST vs AR	min	0.4	3.02	0.61	-0.46	0.83
	mean	9.49	13.05	7.36	8.69	12.57
	max	16.81	19.27	12.5	15.71	20.13
ST vs RF	min	0.17	2.94	0.32	-0.72	2.14
	mean	6.52	10.27	4.5	5.03	7.84
	max	15.3	16.6	9.03	11.12	11.39

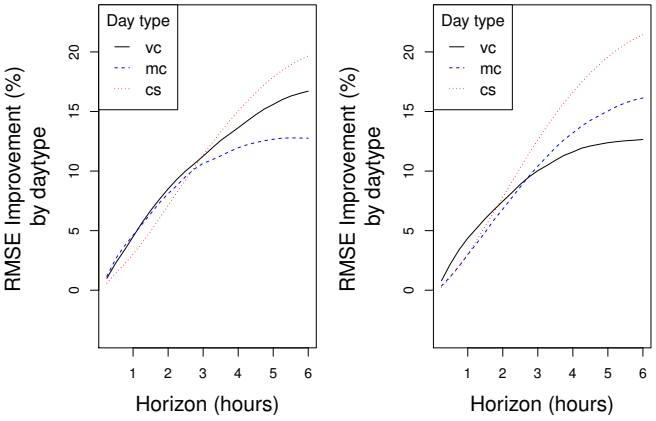

Fig. 6. RMSE Improvement of spatio-temporal model compared to the reference model by day type for two power plants of data set d_1. The day types are very cloudy (vc), moderately cloudy (mc) and clear sky(cs).

The analysis of the performance of the spatio-temporal model can be related to the sky cover. The days of the testing set can be clustered according to sky cover level. We then define three levels of sky coverage: clear sky (cs), moderately cloudy (mc) and very cloudy (vc). These levels were computed using an index based on the ratio of the sum of the daily production to the sum of the simulated irradiation using the ESRA model. Figure 6 presents, for two power plants of d_1, the improvement of the spatio-temporal model compared to the reference model by type of day.

We observe that for the first two hours the improvement on cloudy days exceeds that of clear days. This observation shows that the spatio-temporal model helps to capture the movement of the clouds. The graphs also show that the improvement is greater for clear sky days for the longer horizons and that even on the cloudiest days, the improvement exceeds 5%. This analysis produced similar results for the other power plants.

C. Wind Speed Effect on the Model Performances

We choose the wind speed for the meteorological variable Z as presented in the model extension in part III-C. This choice is motivated by the fact that surface wind speed affects the performance of PV modules given its relation with ambient temperature. Also, wind conditions are generally related to cloud movement, which affects PV production. Note however that by considering surface wind speed, which is in general considerably different from wind speeds in upper layers of the

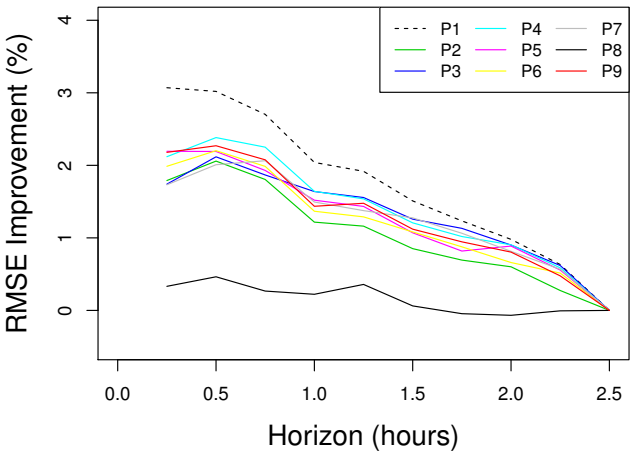

Fig. 7. Data set d_1: RMSE improvement of the spatio-temporal model conditioned by wind speed in comparison with the model without conditioning. Each line represents the improvement obtained for a power plant.

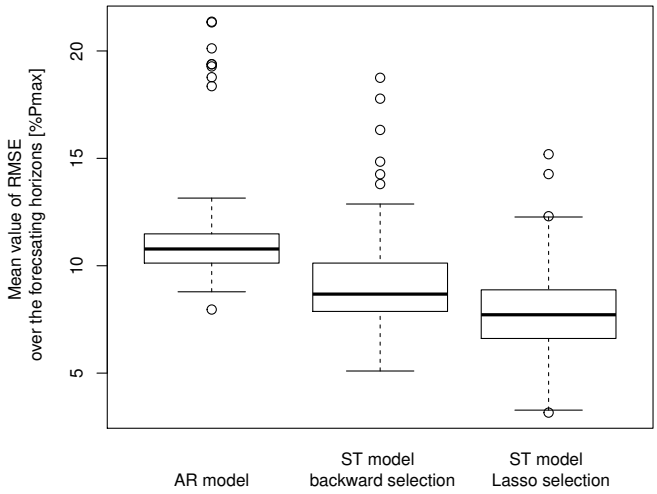

Fig. 8. Data set d_2: Distribution of the mean value (over the 6-hour prediction horizon) of RMSE for the reference model, the spatio-temporal model (ST) with backward selection, and the spatio-temporal model with Lasso selection.

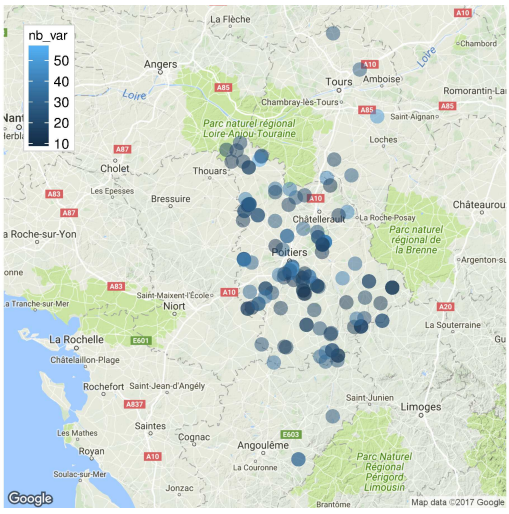

Fig. 9. Data set d_2: Map of the power plants. For each plant, the color is defined by the number of neighboring plants selected by the Lasso.

atmosphere, the aim is not to make an explicit relation with cloud movement.

The spatio-temporal model with a conditioned wind speed parameter was then applied to d_1. Compared to the spatio-temporal model with fixed parameters, this model shows a reduction in RMSE for the first two hours as shown in figure 7. The mean value of this improvement is 2% and the most significant reduction is noted for the first forecasting hour. After two hours, the model with conditioning shows no improvement compared to the model without conditioning. These results are promising and show that there is a potential for improving forecast quality by using adequate meteorological variables within the model.

In the paper [40], the average RMSE improvement of the proposed spatio-temporal model over the reference AR model was about 6% and the maximum RMSE improvement was about 13%. These values are respectively 12% and 20% when applying the spatio-temporal model proposed here.

D. The Variable Selection Contribution: Lasso and AIC

In this section the impact of the different variable selection methods is evaluated. Here we consider the second data set d_2 because the high number of power plants amplifies the dimensionality problem. The spatio-temporal model with the variable selection procedure based on the AIC as described in part III-B was evaluated. The extension of the model with a selection variable procedure based on Lasso regularization (part III-D) was also computed and evaluated on the same data set d_2. Figure 8 represents the dispersion of the mean value (over all prediction horizons) of the RMSE for the reference model and the spatio-temporal model resulting from the two variable selection procedures.

The figure shows that the spatio-temporal model significantly reduces prediction errors compared to the reference model (around 28% reduction in average performance). Moreover, the Lasso variable selection procedure presents lower prediction errors than the selection based on the AIC, showing that the Lasso procedure is more efficient (22% reduction in average performance).

The performance of the Lasso selection variable procedure can also be analyzed by the level of reduction of the dimensionality problem. For each of the power plants of the data set d_2, figure 9 represents the number of neighboring power plants (among the other 135) retained by the Lasso selection. In 75% of cases, the number of variables used is less than 30, while the maximum number used is 57. These numbers show that the Lasso selection variable procedure is quite successful in reducing the dimension of the problem. The results emphasize the interest for the neighboring plants of improving the quality of the PV production forecasts.

V. Conclusion

In this paper we proposed a statistical spatio-temporal model to improve short-term forecasting of photovoltaic production. The non-stationarity issue of the production series was addressed by a new stationarization process. This process demonstrated a clear improvement in terms of forecasting error reduction in comparison with a case in which raw inputs are used. The spatio-temporal model was applied to the stationarized series and showed a significant reduction in forecasting errors compared to regular forecasting techniques. The problem of high dimension data was also addressed by two different variable selection procedures for dimension reduction. The Lasso regularization applied to the spatio-temporal model presents the highest reduction for the forecasts. Moreover, we demonstrate that including the effects of meteorological variables such as wind speed in the spatio-temporal results in an additional reduction of the forecasting error level of PV production.

Further work could investigate beyond the linear modeling of the spatio-temporal data using more complex relations like polynomial estimations or splines. The integration of meteorological data could also be investigated, either as a parameter of the coefficient estimated in the spatio-temporal model, or by integrating sky images obtained by cameras or satellites. A probabilistic model that uses information on geographically distributed power plants to produce forecasts could also be investigated.

Acknowledgment

The authors would like to thank the French industrials Coruscant and Hespul for providing the PV data. They would also like to thank Prof. Philippe Blanc (MINES ParisTech) for his advice on solar radiation data treatment and clear sky models and for providing ESRA data.

References

[1] C. W. Potter, A. Archambault, and K. Westrick, "Building a smarter smart grid through better renewable energy information," in *Proceedings of Power Systems Conference and Exposition*, Seattle, WA, USA, March 2009.

[2] P. Pinson, C. Chevallier, and G. N. Kariniotakis, "Trading Wind Generation From Short-Term Probabilistic Forecasts of Wind Power," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1148–1156, Aug. 2007.

[3] X. Wu, X. Hu, S. Moura, X. Yin, and V. Pickert, "Stochastic control of smart home energy management with plug-in electric vehicle battery energy storage and photovoltaic array," *Journal of Power Sources*, vol. 333, pp. 203–212, Nov. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037877531631357X

[4] R. H. Inman, H. T. C. Pedro, and C. F. M. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in Energy and Combustion Science*, vol. 39, no. 6, pp. 535–576, Dec. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360128513000294

[5] V. Kostylev and A. Pavlovski, "Solar power forecasting performances - towards industry standards," in *Proceedings of 1st International Workshop on the Integration of Solar Power into Power Systems*, Aarhus, Denmark, October 2011.

[6] J. Shi, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting power output of photovoltaic system based on weather classification and support vector machine," in *Industry Applications Society Annual Meeting (IAS), 2011 IEEE*, Oct 2011, pp. 1–6.

[7] J. G. da Silva Fonseca, T. Oozeki, T. Takashima, G. Koshimizu, Y. Uchida, and K. Ogimoto, "Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in kitakyushu, Japan," *Progress in Photovoltaics: Research and Applications*, vol. 20, no. 7, pp. 874–882, 2012. [Online]. Available: http://dx.doi.org/10.1002/pip.1152

[8] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, "Predicting solar generation from weather forecasts using machine learning," in *Proceedings of IEEE International Conference on Smart Grid Communications, (IEEE SmartGridComm)*, Brussels, Belgium, October 2011. [Online]. Available: http://ieeexplore.ieee.org/

[9] O. Perpin and E. Lorenzo, "Analysis and synthesis of the variability of irradiance and {PV} power time series with the wavelet transform," *Solar Energy*, vol. 85, no. 1, pp. 188 – 197, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X10002811

[10] M. Zamo, O. Mestre, P. Arbogast, and 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*, vol. 105, pp. 792 – 803, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X13005239

[11] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Solar Energy*, vol. 83, no. 10, pp. 1772 – 1783, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X09001364

[12] P. Bacher, H. Madsen, B. Perers, and H. A. Nielsen, "A non-parametric method for correction of global radiation observations," *Solar Energy*, vol. 88, pp. 13 – 22, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X12003891

[13] H. T. Pedro and C. F. Coimbra, "Assessment of forecasting techniques for solar power production with no exogenous inputs," *Solar Energy*, vol. 86, no. 7, pp. 2017 – 2028, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X12001429

[14] C. Monteiro, T. Santos, L. A. Fernandez-Jimenez, I. J. Ramirez-Rosado, and M. S. Terreros-Olarte, "Short-term power forecasting model for photovoltaic plants based on historical similarity," *Energies*, vol. 6, no. 5, p. 2624, 2013. [Online]. Available: http://www.mdpi.com/1996-1073/6/5/2624

[15] V. G. Berdugo, C. Chaussin, and L. D. et al., "Analog method for collaborative very-short-term forecasting of power generation from photovoltaic systems," Patent 1 154 438.

[16] E. Lorenz, D. Heinemann, and C. Kurz, "Local and regional photovoltaic power prediction for large scale grid integration: Assessment of a new algorithm for snow detection," *Progress in Photovoltaics: Research and Applications*, vol. 20, no. 6, pp. 760–769, 2012. [Online]. Available: http://dx.doi.org/10.1002/pip.1224

[17] Y. A., S. T., and S. A. et al., "Application of neural network to one-day-ahead 24 hours generating power forecasting for photovoltaic system," in *Proceedings of the International Conference on Intelligent Systems Applications to Power Systems*, Kaohsiung, Taiwan, November 2007.

[18] Y. HUANG, J. LU, and C. L. et al., "Comparative study of power forecasting methods for pv stations," in *Proceedings of the International Conference on Power System Technology*, Zhejiang, Zhejiang, China, October 2010.

[19] C. Tao, D. Shanxu, and C. Changsong, "Forecasting power output for grid-connected photovoltaic power system without using solar radiation measurement," in *Proceedings of the IEEE International Symposium on Power Electronics for Distributed Generation Systems*, Hefei, China, June 2010. [Online]. Available: http://ieeexplore.ieee.org/

[20] L. A. Fernandez-Jimenez, A. Muoz-Jimenez, A. Falces, M. Mendoza-Villena, E. Garcia-Garrido, P. M. Lara-Santillan, E. Zorzano-Alba, and P. J. Zorzano-Santamaria, "Short-term power forecasting system for photovoltaic plants," *Renewable Energy*, vol. 44, pp. 311 – 317, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148112001516

[21] A. Mellit and A. M. Pavan, "A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected {PV} plant at trieste, italy," *Solar Energy*, vol. 84, no. 5, pp. 807 – 821, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X10000782

[22] M. Khn, C. Juhlin, H. Held, V. Bruckman, T. Tambach, T. Kempka, S. Jerez, R. Trigo, A. Sarsa, R. Lorente-Plazas, D. Pozo-Vzquez, and J. Montvez, "European geosciences union general assembly 2013, egudivision energy, resources & the environment, ere spatio-temporal complementarity between solar and wind power in the iberian peninsula," *Energy*

Procedia, vol. 40, pp. 48 – 57, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1876610213016019
[23] J. Dowell, S. Weiss, D. Hill, and D. Infield, "Short-term spatio-temporal prediction of wind speed and direction," *Wind Energy*, vol. 17, no. 12, pp. 1945–1955, 2014. [Online]. Available: http://dx.doi.org/10.1002/we.1682
[24] J. Tastu, P. Pinson, E. Kotwa, H. Madsen, and H. A. Nielsen, "Spatio-temporal analysis and modeling of short-term wind power forecast errors," *Wind Energy*, vol. 14, no. 1, pp. 43–60, 2011. [Online]. Available: http://dx.doi.org/10.1002/we.401
[25] R. Girard and D. Allard, "Spatio-temporal propagation of wind power prediction errors," *Wind Energy*, vol. 16, no. 7, pp. 999–1012, 2013. [Online]. Available: http://dx.doi.org/10.1002/we.1527
[26] M. He, L. Yang, J. Zhang, and V. Vittal, "A Spatio-Temporal Analysis Approach for Short-Term Forecast of Wind Farm Generation," *IEEE Transactions on Power Systems*, vol. 29, no. 4, pp. 1611–1622, Jul. 2014.
[27] J. Tastu, P. Pinson, P. J. Trombe, and H. Madsen, "Probabilistic Forecasts of Wind Power Generation Accounting for Geographically Dispersed Information," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 480–489, Jan. 2014.
[28] M. Sherman, *Spatial Statistics and Spatio-Temporal Data*. Wiley, 2011.
[29] J. Bosch and J. Kleissl, "Cloud motion vectors from a network of ground sensors in a solar power plant," *Solar Energy*, vol. 95, pp. 13 – 20, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X13002193
[30] M. Lave and J. Kleissl, "Cloud speed impact on solar variability scaling application to the wavelet variability model," *Solar Energy*, vol. 91, pp. 11 – 21, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X13000406
[31] S. Quesada-Ruiz, Y. Chu, J. Tovar-Pescador, H. Pedro, and C. Coimbra, "Cloud-tracking methodology for intra-hour {DNI} forecasting," *Solar Energy*, vol. 102, pp. 267 – 275, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X14000486
[32] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker Jr., D. Renn, and T. E. Hoff, "Validation of short and medium term operational solar radiation forecasts in the US," *Solar Energy*, vol. 84, no. 12, pp. 2161–2172, Dec. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X10002823
[33] C. A. Glasbey and D. J. Allcroft, "A Spatiotemporal Auto-Regressive Moving Average Model for Solar Radiation," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 57, no. 3, pp. 343–355, 2008. [Online]. Available: http://www.jstor.org/stable/20492608
[34] D. Yang, C. Gu, Z. Dong, P. Jirutitijaroen, N. Chen, and W. M. Walsh, "Solar irradiance forecasting using spatial-temporal covariance structures and time-forward kriging," *Renewable Energy*, vol. 60, pp. 235–245, Dec. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148113002759
[35] A. Tascikaraoglu, B. Sanandaji, G. Chicco, V. Cocina, F. Spertino, O. Erdinc, N. Paterakis, and J. P. Catalao, "Compressive Spatio-Temporal Forecasting of Meteorological Quantities and Photovoltaic Power," *IEEE Transactions on Sustainable Energy*, vol. PP, no. 99, pp. 1–1, 2016.
[36] R. Dambreville, P. Blanc, J. Chanussot, and D. Boldo, "Very short term forecasting of the Global Horizontal Irradiance using a spatio-temporal autoregressive model," *Renewable Energy*, vol. 72, pp. 291–300, Dec. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S096014811400398X
[37] V. P. A. Lonij, A. E. Brooks, A. D. Cronin, M. Leuthold, and K. Koch, "Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors," *Solar Energy*, vol. 97, pp. 58–66, Nov. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X13003125
[38] J. D. Patrick, J. L. Harvill, and C. W. Hansen, "A semiparametric spatio-temporal model for solar irradiance data," *Renewable Energy*, vol. 87, Part 1, pp. 15–30, Mar. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148115303542
[39] C. Yang, A. A. Thatte, and L. Xie, "Multitime-Scale Data-Driven Spatio-Temporal Forecast of Photovoltaic Generation," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 1, pp. 104–112, Jan. 2015.
[40] X. G. Agoua, R. Girard, and G. Kariniotakis, "Spatio-temporal models for photovoltaic power short-term forecasting," in *Solar Integration workshop 2015*, Brussels, Belgium, Oct. 2015. [Online]. Available: https://hal-mines-paristech.archives-ouvertes.fr/hal-01220321
[41] R. J. Bessa, A. Trindade, C. S. P. Silva, and V. Miranda, "Probabilistic solar power forecasting in smart grids using distributed information," *International Journal of Electrical Power & Energy Systems*, vol. 72, pp. 16–23, Nov. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0142061515000897
[42] H. C. Hottel, "A simple model for estimating the transmittance of direct solar radiation through clear atmospheres," *Solar Energy*, vol. 18, pp. 129 – 134, 1976.
[43] C. Rigollier, O. Bauer, and L. Wald, "On the clear sky model of the ESRA European Solar Radiation Atlas with respect to the heliosat method," *Solar Energy*, vol. 68, no. 1, pp. 33–48, Jan. 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X99000559
[44] N. A. Engerer and F. P. Mills, "KPV: A clear-sky index for photovoltaics," *Solar Energy*, vol. 105, pp. 679–693, Jul. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X14002151
[45] T. Robert, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.
[46] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

Xwégnon Ghislain Agoua engineer graduate of ENSAI (Ecole nationale de la statistique et de l'analyse de l'information), France, in 2014.
He is currently a PhD student at MINES ParisTech, PSL - Research University, PERSEE - Centre for Processes, Renewable Energies and Energy Systems. He is mostly interested in statistical modeling, forecasting techniques, time series analysis, spatio-temporal regression models, and their applications to photovoltaic generation modeling.

Robin Girard received a Master's degree (2004) in Computer Science and Applied Mathematics from INPG in Grenoble, France and a PhD degree (2008)in applied Mathematics from Joseph Fourier University in Grenoble.
He is currently a Research Engineer at the Centre of Energy and Processes of the Ecole des Mines de Paris. His research interests include wind and solar power forecast, optimization in planning of energy production and spatio-temporal patterns of renewable power production.

George Kariniotakis (S95-M02-SM11) was born in Athens, Greece. He received his Eng. and M.Sc. degrees from Greece in 1990 and 1992 respectively, and his Ph.D. degree from Ecole des Mines de Paris in 1996. He is currently with the Centre PERSEE of MINES ParisTech as a senior scientist and head of the Renewable Energies and Smartgrids Group. He has authored more than 200 scientific publications in journals and conferences. He has been involved as participant or coordinator in more than 40 R&D projects in the fields of renewable energies and distributed generation. Among them, he was the coordinator of some major EU projects in the field of wind power forecasting such as Anemos, Anemos.plus and SafeWind projects. His scientific interests include among others timeseries forecasting, decision making under uncertainty, modelling, management and planning of power systems.