



HAL
open science

Probabilistic Model for Spatio-Temporal Photovoltaic Power Forecasting

Xwégnon Ghislain Agoua, Robin Girard, Georges Kariniotakis

► **To cite this version:**

Xwégnon Ghislain Agoua, Robin Girard, Georges Kariniotakis. Probabilistic Model for Spatio-Temporal Photovoltaic Power Forecasting. *IEEE Transactions on Sustainable Energy*, 2019, 10 (2), pp.780 - 789. 10.1109/TSTE.2018.2847558 . hal-01817629

HAL Id: hal-01817629

<https://hal.science/hal-01817629>

Submitted on 18 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Probabilistic Model for Spatio-Temporal Photovoltaic Power Forecasting

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

Abstract—Photovoltaic (PV) power generation is characterized by significant variability. Accurate PV forecasts are a prerequisite to securely and economically operating electricity networks, especially in the case of large-scale penetration. In this paper, we propose a probabilistic spatio-temporal model for the PV power production that exploits production information from neighboring plants. The model provides the complete future probability density function of PV production for very short-term horizons (0-6 hours). The method is based on quantile regression and a L_1 penalization technique for automatic selection of the input variables. The proposed modeling chain is simple, making the model fast and scalable to direct on-line application. The performance of the proposed approach is evaluated using a real-world test case, with a high number of geographically distributed PV installations and by comparison with state-of-the-art probabilistic methods.

Index Terms—Lasso, photovoltaic generation, probabilistic forecasts, quantile regression, reliability, sharpness, spatio-temporal

I. INTRODUCTION

RENEWABLE Energy Sources (RES), including photovoltaic (PV) production, are being developed in many countries as a response to the need for clean energy solutions. PV production is characterized by high variability and uncertainty due to its inherent dependence on changing meteorological conditions. Variability and uncertainty are a challenge for network operators especially in systems with large-scale RES integration. Short-term forecasting of PV production is a prerequisite for the economic and secure management of power systems, reduction of reserve costs [1] and market participation of PV producers. It is also important to ensure the competitiveness of renewable energy technologies [2].

The literature features several methods to forecast PV production. Detailed reviews of the state-of-the-art are provided in [3]–[5]. They can be classified according to the forecast horizon, the available data, and the type of approach, which may be based on statistics, physics or a hybrid combination [4], [6]. Although early methods were deterministic, probabilistic approaches are increasingly popular since they provide additional information about the distribution of future production

This work was carried out within the research project entitled “Improvement of PV power forecasting and predictive management including storage solutions”, supported by the company Corusant SA in the frame of its participation to a tender of the French Energy Regulator CRE for the development of PV plants above 250 kWc.

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, robin.girard, georges.kariniotakis each with @mines-paristech.fr)

and uncertainty in the forecasts. Some of these probabilistic approaches are based on Numerical Weather Prediction (NWP) models or sky imaging and provide ensemble forecasts of the future PV generation [7]–[9]. Analog ensembles [10], regression trees [11], [12] and k-nearest neighbors (kNN) [13]–[15] are also found in the related literature on probabilistic PV forecasting. A wide range of Artificial Neural Networks (ANN) based models also exist for short-term PV power production [16]–[21]. These models have evolved from simple neural networks, to radial neural networks (more suitable for time series prediction) and more recently to deep learning methods [22], [23].

Recently the focus has shifted to improving short-term predictability of solar irradiance or PV production by considering off-site information as input to the models. Such models are commonly referred to as spatio-temporal models. Models for solar radiation forecasting are based on autoregressive models [24]–[26], geostatistical models [27]–[29] or classification models [30]. Physical models combined with NWP forecasts [31], [32] and semi-parametric models [33] can also be used for spatio-temporal forecasting of solar radiation. ANN-based models are also used for spatio-temporal solar power forecasting [34]–[36]. For the case of PV power forecasting, most spatio-temporal models use off-site meteorological measurements and neighboring site information to improve the forecast for the site of interest. These models are autoregressive models [37]–[39] or classification models [30]. The classification model proposed in [40] handles data acquisition and privacy issues.

The spatial-temporal models listed above provide deterministic forecasts. In contrast to day-ahead forecasting, very few spatial-temporal models for short-term probabilistic forecasting feature in the literature. These models are based on regression trees [12], the kNN method [13], the combination of a vectorial autoregressive model and gradient boosting [41], multivariate predictive distributions, [42] and Gaussian random fields [43].

One of the biggest challenges in spatio-temporal forecasting is to propose models that are able to handle the dimensionality issues raised by the large amount of explanatory variables. This is a prerequisite for the models so that they can generalize efficiently when they “see” new data. In this paper, we propose a short-term probabilistic forecasting methodology that exploits information available from a large number of PV installations and includes such a process for the automated selection of explanatory variables. The aim is to improve predictability in the time range of 0-6 hours.

In a previous paper [44], we proposed a deterministic

spatio-temporal model for PV power forecasting. This paper makes several additional contributions, compared to both the previous paper and to the related literature on probabilistic PV forecasting. These contributions are: 1) the use of off-site data (neighboring production data) to generate probabilistic forecasts with improved properties and performance compared to the standard approach without off-site data 2) the introduction of an automated variable selection procedure for the PV power forecasts inside the model, which avoids the dimensionality problem 3) the integration of NWP forecasts as explanatory variables despite the short-term frame; this allows us to assess the effect of the NWP variables on the probabilistic forecasts. In probabilistic models in the literature [35], [41], [45] the variable selection is done either empirically or with the variable importance criteria for decision trees; 4) two probabilistic models chosen from the most efficient ones in the literature are used as references for our model evaluation.

In contrast to methods proposed in the literature for this time range, our approach is not based on information from sky cameras or satellite images. It uses available data from off-site PV plants. The model uses the transfer of spatial and temporal information within the network of power plants to capture the meteorological perturbations. Unlike in [46], these spatial and temporal correlations between power plants are useful for short-term horizons as they offer good spatial and temporal resolution.

Moreover, in the proposed model we use NWP forecasts as additional explanatory input, which is not standard practice in the literature for models developed especially for this time scale. NWPs are a standard input for models that address day-ahead or longer horizons. Such models also cover horizons shorter than 6h. It is however widely recognized in renewable generation forecasting that, for the first hours, NWPs do not obtain performance improvements w.r.t. persistence. To achieve such improvements, it is necessary to include recent measurements, such as input. This observation, coupled with the need for frequent forecast updates (i.e. every 30 minutes) for applications like intraday trading, have motivated the development of purely statistical models based only on measurements. The standard spatiotemporal approach comes into this category. However, we have observed that including NWPs in this family of models can provide information on weather condition trends for horizons around 6 hours ahead. In cases like wind prediction, this inclusion brings improvements w.r.t. persistence, amounting to double the improvements without NWPs for 6 hours ahead [3]. This inclusion of NWPs in the family of short-term models that operate with very frequent updates goes against the general tendency in the literature. However, this approach allows for better coupling between very short-term spatial-temporal forecasts and standard day-ahead forecasts since NWPs reduce the mismatch between two types of forecast that can be confusing for end-users. The most relevant NWP forecasts were selected by measuring their dependence on production data.

The generation of the probabilistic forecast was done using a quantile regression scheme. Therefore the variable selection process using L1-penalization that we propose in this paper is different from the one used in [44] as it is done in a L1-

optimization model (quantile regression). We have also added the constraints of different parameters for each quantile and each time step.

This paper puts forward the non-stationary characteristic of PV power series. The spatio-temporal approach exploits correlations in the data from geographically dispersed sites. It is thus necessary to efficiently eliminate the deterministic effect that is related to the course of the sun. This issue has been treated by applying an advanced stationarization process to the production series as proposed in [44]. The evaluation is carried out on a real world test case of 185 PV installations in France. This test case is a good illustration of a system with high penetration of renewable energy and also allows us to demonstrate the full potential of a spatio-temporal approach for probabilistic forecasting when considering a highly dimensional problem.

The paper is structured as follows: the data and an analysis of the spatio-temporal correlations are presented in section II. The proposed spatio-temporal probabilistic models are presented in section III. The evaluation and analysis of the performances of the forecasts are treated in section IV. Finally, the conclusions of the study are discussed in section V.

II. THE INTEREST OF SPATIO-TEMPORAL MODELING

The aim of this section is to illustrate the interest and potential of spatio-temporal modeling in short-term PV power forecasting. A real-world case study is introduced followed by an analysis illustrating that useful information can be extracted from geographically distributed PV installations.

A. Data

The data set considered here, denoted as d , comprises the output of 905 PV power inverters in a mid-west region of France with peak power ranging from 3.2 kWc to 58 kWc. This number 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 with a 15 min temporal resolution. The locations of the power plants are represented in Figure 1. After data cleaning, 136 power plants were retained. In the following, the power plants are labeled $P_{i,1 \leq i \leq 136}$. The production series have been stationarized employing the same procedure as that proposed in [44], which is used thereafter.

B. Spatial-Temporal Correlation

The existence of a spatial-temporal pattern is evaluated by an analysis of the correlation between the stationarized production series. In these series, the effects of the sun position on cross-correlations are absent.

Figure 2 presents the histograms of the lagged cross-correlation values for three classes of distance. For a couple of stationarized series s_1, s_2 for example, the value of the correlation is obtained by evaluating the correlation between s_1 and the lagged series $lag(s_2, k)$. The maximum lag used is the maximum horizon considered for the forecasts, here

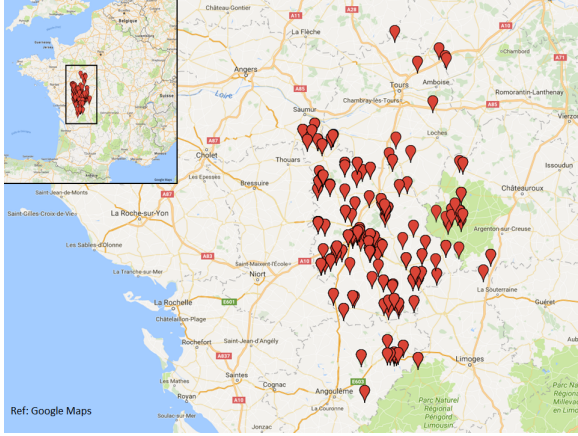


Fig. 1. The power plants of the test case d .

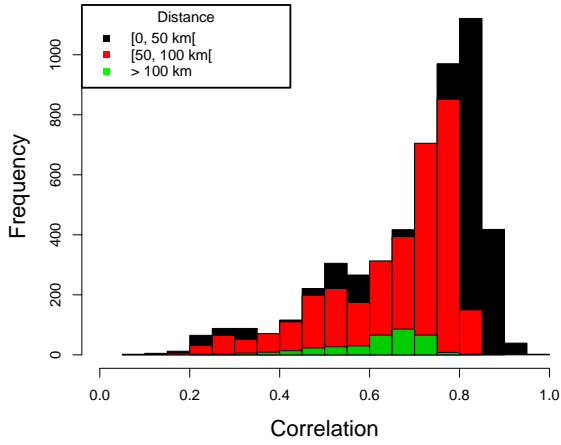


Fig. 2. Histogram by class of distance of the cross-correlation between the lagged production series after stationarization.

6 hours. The same evaluation is made by switching s_1 and s_2 . The value retained is the maximum value. The correlation values are quite high, concentrated in the interval $[0.4 - 0.8]$ and, decrease with distance. Therefore, we can assume that these significant correlation values describe a spatial transfer of information between the power plants. This transfer, mainly due to cloud movements, can be used to anticipate meteorological perturbations and to improve the forecasts using production data only.

From a temporal point of view, Figure 3 presents the time lags for which the cross correlations described above reach their maximum. The idea behind this analysis is to determine the temporal limit of the propagation of correlation between the power plants. The figure shows that in the majority of cases, the maximum correlation is obtained for 15 minutes, but for some power plant couples it can be reached for 4 hours or more. The models proposed in the next section aim to exploit this spatial and temporal information.

III. PROBABILISTIC MODELS FOR SPATIAL-TEMPORAL PV FORECASTING

We consider here two models chosen from the most efficient probabilistic forecasting methods in the literature as reference

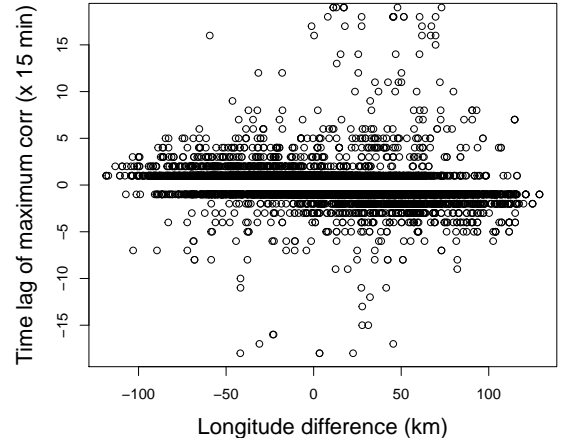


Fig. 3. Time lags for which the correlation between each power plant couple is at a maximum. One point per couple.

models: Kernel Density Estimation (KDE) and the Extreme Learning Machine (ELM) method proposed in [47]. The KDE model is a well-known method used to evaluate the future density of PV power [48], [49]. An information-based criterion is proposed to select the input variables. These two models respectively use only “local” on-site information from the PV installation the power of which is predicted. These models can thus be used to evaluate the contribution of the proposed spatial-temporal probabilistic approach. The advanced method we propose is based on the quantile regression approach, which we adapted for high dimensionality problems. We describe below the principle of these methods, the conditions of their implementation and the modifications proposed to improve their efficiency.

A. Kernel Density Estimation

KDE is a non-parametric approach for density estimation [50]–[52]. This method reduces the level of estimation errors compared to parametric approaches as there is no hypothesis on the underlying distribution. The minimization problem consists in providing an estimation of the density f of a random variable X . The n -dimensional multivariate kernel-estimator expression is:

$$\hat{f}(x) = \frac{1}{N|\mathbf{H}|} \sum_{i=1}^N K(\mathbf{H}^{-1}(x - x_i)) \quad (1)$$

where x is the point where the evaluation is made, $x_i, i = 1 \dots N$ the data. \mathbf{H} is a $n \times n$ matrix called the bandwidth matrix, which controls the smoothing and $|\mathbf{H}|$ its determinant. K is the kernel function. The kernel function K and the bandwidth matrix \mathbf{H} are the two parameters which have to be determined to apply the KDE. The impact of the kernel function on the estimation quality is low and a standard Gaussian kernel implies a high computation cost [53]. In this paper, we have chosen to use the Epanechnikov kernel function:

$$K(u) = \frac{3}{4\sqrt{5}}(1 - u^2/5) \quad u \in [-1, 1]. \quad (2)$$

The multivariate version of this kernel function is the product $K(u) = \prod_{j=1}^n K(u_j)$. The smoothing matrix \mathbf{H} is the most important parameter as it has a strong influence on the quality of the estimation [52]. Several approaches are possible for selecting the appropriate bandwidth matrix [54], [55]. In this paper, we use the smoothed cross validation technique [56]. Moreover, as the PV production data are positive and bounded, we adapt the general method to our needs by using boundary correction as proposed in the literature [52], [53].

Let $Y \in \mathbb{R}^p$ be the random variable, the realizations of which are the production of a power plant and $X \in \mathbb{R}^p$ the explanatory variables. The objective is to compute the probability density function of the conditioned random variable $Y_{t+k}|X_t$ where t is the time the prediction is made and k the horizon. This density function is obtained by:

$$f_{Y_{t+k}|X_t} = \frac{f_{Y_{t+k}, X_t}}{f_{X_t}} \quad (3)$$

The forecast probability density function is :

$$\hat{f}_{Y_{t+k}|X_t} = \frac{1}{|\mathbf{H}|} \sum_{i=1}^N w(x, x_i) K(\mathbf{H}^{-1}(y - y_i)) \quad (4)$$

where

$$w(x, x_i) = \frac{K(\mathbf{H}^{-1}(x - \mathbf{x}_i))}{\sum_{j=1}^N K(\mathbf{H}^{-1}(x - \mathbf{x}_j))}. \quad (5)$$

The mutual information criterion [57] has been used for choosing the inputs of the KDE model. It is a measure of "distance" (Kullbac-Leiber) between two probability density functions. Here this criterion is used to evaluate the distance between the probability density functions of the PV production and those of the meteorological variables. This evaluation is carried out for each meteorological variable. The values of the information criterion are used to classify the meteorological variables according to their proximity to the probability density of the production, following which the most relevant inputs are selected for the KDE model. Let X be the random variable representing a meteorological variable and Y that of the PV production. If f_X and f_Y denote their respective probability density functions, the average mutual information between Y and X is:

$$I(X, Y) = \int_Y \int_X f_{X,Y} \log \left(\frac{f_{X,Y}}{f_X \cdot f_Y} \right) dx dy. \quad (6)$$

A null value of this criterion indicates that the variables X and Y are independent.

B. ELM-Based Estimation

The Extreme Learning Machine method is a non-parametric density forecasting method presented and developed in [47]. It is based on feed forward neural networks but does not require tuning for the hidden layer parameters (the parameters are randomly selected making it fast). The proposed configuration of the ELM-based estimation model for this paper is the following:

- one model is tuned for each power plant;

- the predictors are the past solar observations of the site of interest with respective time lags and the NWP variables selected by the mutual information criterion presented in subsection III-A;
- the results are generated as predicted quantiles for a 6-hour horizon with 15-min time-steps; the forecasting model is updated monthly
- the non-crossing quantile constraint has been applied.

C. The proposed model: the QR-Lasso

The aim of quantile regression [58] is to provide an estimation of the following cumulative distribution function:

$$F(y|X = x) = \mathbb{P}(Y \leq y|X = x) = \mathbb{E}(\mathbf{1}_{Y \leq y}|X = x). \quad (7)$$

This estimation is made by providing information about a significant number of points describing the distribution known as the quantiles. For a continuous probability distribution function, the quantile of level $\alpha \in (0, 1)$, $\hat{q}_\alpha(Y|X)$ is

$$q_\alpha(Y|X) = \mathbb{F}^{-1}(\alpha) = \inf\{y, \mathbb{F}(y|X = x) \geq \alpha\}. \quad (8)$$

This quantile is a solution of the minimization problem:

$$q_\alpha(Y) = \arg \min_g \mathbb{E}[\rho_\alpha(Y - g(X))] \quad (9)$$

where $\rho_\alpha(u) = u(\alpha - \mathbf{1}_{\{u < 0\}})$ is the loss function called the pinball loss. The problem (9) can be generalized to the conditional quantile of level α as :

$$q_\alpha(Y|X) = \arg \min_g \mathbb{E}[\rho_\alpha(Y - g(X))|X = x]. \quad (10)$$

When assuming a linear relation between Y and X , the estimation can be made on the observation set using:

$$\hat{q}_\alpha = \arg \min_\beta \sum_{i=1}^n \rho_\alpha(Y_i - X_i' \beta). \quad (11)$$

The equation (11) does not have an explicit solution; the solution has been found numerically, despite the fact that the pinball loss is neither differentiable in 0 nor strictly convex. The problem is then transformed into a linear optimization problem solved by the simplex algorithm (for small samples) or interior point method (for large samples) [59]. The condition of non-crossing quantiles has been taken into account. In practice, this condition is implemented by adding some constraints to the problem that specify that each quantile value is positive, lower than the next value, and lower than the maximum.

D. Adaptation for the High Dimensional Input

Spatio-temporal forecasting, when considering numerous PV plants, raises the question of choosing the appropriate input variables for the models involved. Possible inputs include the production series of the predicted site and the neighboring sites, and the lagged version of these series. The NWP meteorological data can also be integrated into the models. Given that the PV plants may cover a large area of tens of kilometers, several points of the NWP grid can be considered. The above information represents a high amount of potential inputs for the forecasting models and raises the issue of their

dimensionality and over-fitting. We propose here a variable selection process, which we have adapted to the quantile regression.

The method is based on a direct L1-penalization called the Least Absolute Shrinkage and Selection Operator (LASSO) [60]. This penalization is applied directly to the estimation process of the quantile regression model. The estimator is defined for the quantile of level α as:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \{ \mathbb{E}[\rho_{\alpha}(Y - X'\beta)] + \lambda \|\beta\|_1 \}. \quad (12)$$

This penalization also helps to avoid over-fitting in the model. For values of the penalization parameter λ that are high enough, some coefficients β are set to zero, thus allowing a selection to be made between the input variables.

The problem (12) can be reformulated as

$$\underset{\beta}{\operatorname{argmin}} \{ \mathbb{E}[\rho_{\alpha}(Y - X'\beta)] \} \quad \text{st} \quad \|\beta\|_1 \leq \hat{R}_{\lambda}. \quad (13)$$

The equations (12) and (13) are equivalent. The latter is the most common form of optimization problem. The Lasso is also more suitable for operational models as its run-time is lower than those of other methods, like L2-penalization, gradient boosting or the stepwise process. The L1 loss ensures a variable selection that makes the model perform well with a reasonable number of variables and is stricter than the L2 loss. More details about the regularization paths can be found in the literature [60].

This modification of the standard quantile regression proposed here is essential in the spatial-temporal paradigm due to the high number of potential inputs. The new quantile regression model including this LASSO variable selection procedure is hereafter called QR-Lasso. The performance of this model is evaluated against the reference KDE model.

IV. EVALUATION OF THE FORECASTS

The performance of the proposed model is evaluated for horizons up to 6 hours ahead with a 15-min time step. It is considered that forecasts are updated every 15 minutes. This performance is compared to the reference model. The NWP data were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF). These forecasts are updated 4 times a day and are given with a 3-hourly temporal resolution. 15-min values were obtained through interpolation. The models were developed using the software R [61] with the packages *quantreg*, *glmnet*, *elmNN* and some specific functions developed for this study (matrice estimations for KDE, penalized quantile, etc). The probabilistic (distribution) forecasts will be characterized by their quantiles. The last third of the data set (around 5 months) was used as a testing set. The QR-Lasso model is updated for each time step for each decile. A single update step took approximately 5 seconds (on a 12 GB RAM computer).

A. Variables Selection Through Mutual Information

The variable selection procedure allows us to choose the relevant variables from numerous available inputs, measurements and NWPs, for the case where a large number of PV plants are involved in the spatio-temporal modeling.

TABLE I
NORMALIZED MUTUAL INFORMATION VALUES FOR 4 PV PLANTS OF THE DATA SET COMPUTED BETWEEN PV PRODUCTION AND NWP VARIABLES.

NWP Variables	Mutual information (%)			
	P_1	P_2	P_3	P_4
Temperature (T)	72.38	70.74	80.79	78.22
Relative Humidity (RH)	70.56	74.56	74.83	71.27
Wind Direction (WD)	25.11	21.88	26.11	21.86
Wind Speed (WS)	6.18	5.78	6.57	5.46
Precipitation	0.18	0.48	0.26	0.27

In Table I, we present the normalized mutual information between the PV production of four PV plants and the NWP variables. The mutual information has not been computed for the top net solar radiation variable since this variable is the main driver of the PV production process. The table shows that the two most important variables are the temperature and the humidity followed by the wind direction. The importance of the precipitation level is the lowest.

Using this process, we select the following input NWP variables for the reference KDE model, which will then be multivariate:

- the top net solar radiation (TSR),
- the temperature (T),
- the relative humidity (RH),
- the wind direction (WD).

B. Analysis of the Lasso Variable Selection

The potential input of the proposed spatio-temporal QR-Lasso model for each PV plant includes the measured power of the other 135 plants, the lagged values of these measurements and finally the NWP variables corresponding to the 4 surrounding NWP grid points. The subset of NWP variables selected by the mutual information criterion is considered. Then for each PV plant $P_i, i = 1, \dots, 136$, the number of total potential inputs is $819 = 4$ (NWP forecasts) + $(136 \times 6) - 1$ (6 lags). Figure 4 presents the number of non-null coefficients related to PV power inputs selected by the QR-Lasso model for each decile. The maximum number of variables selected is 320, showing that the selection process is efficient. The performance of the variable selection process is also evaluated by examining the geographical position of the selected PV plants with respect to the position of the PV plant of interest. Figure 5 presents the positions of the selected PV plants when forecasting the median of the production of two PV plants for a 6-hour horizon. The numbers of PV plants selected from the 136 initial plants are 33 (left) and 43 (right). For the two power plants of interest presented, the choice of neighboring plants is quite homogeneous. Some power plants close to the plant of interest do not provide more knowledge to the model (in terms of error reduction) than the selected ones. To ensure the parsimony, these plants are thus not selected by the model.

C. The Performance of the QR-Lasso Spatial-Temporal Approach

The quality of the probabilistic forecasts produced with the selected variables can be evaluated with several criteria [62].

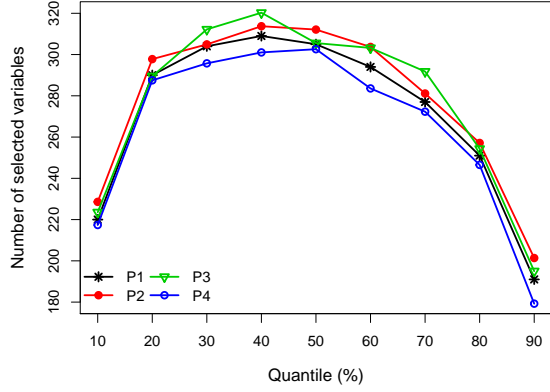


Fig. 4. Number of non-null coefficients computed by the QR-Lasso model by decile level (the meteorological coefficients are not included). Four PV plants ($P_1 - P_4$) from the test case are represented and each line corresponds to one power plant. The horizon is 6 hours.

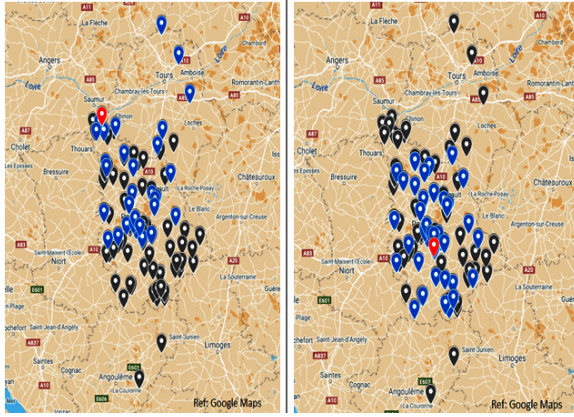


Fig. 5. The power plants selected by the QR-Lasso model (in blue) when forecasting the median of the future production for a 6-hour horizon for two PV plants (in red).

Here we limit our study to the two most common criteria, i.e. reliability and sharpness, and a combined criterion that represents overall skill.

Reliability is a criterion that describes the ability of the probabilistic forecast model to match the expected observation frequencies. For \hat{q}^α a predicted quantile of the level α , the reliability criterion verifies whether the associated proportion of data observed under this quantile is equal to the expected α . To evaluate the reliability of the full predictive density for each of our models, we evaluate the reliability for quantile q^{10} to q^{90} with an increasing step of 10. The quantiles can be interpreted as estimated coverage rates, which then allows us to evaluate the reliability through increasing coverage rate analysis. Formally, reliability can be defined for the quantile level (or coverage rate) α and horizon h as:

$$Rel_h^\alpha = \alpha - \frac{1}{N} \sum_{t=1}^N \mathbf{1}\{y_{t+h|t} \leq \hat{q}_{t+h|t}^\alpha\} \quad (14)$$

where t is the time when the prediction is made.

Figure 6 presents for these quantiles (predictive intervals) the observed deviation from the nominal coverage for 3h hori-

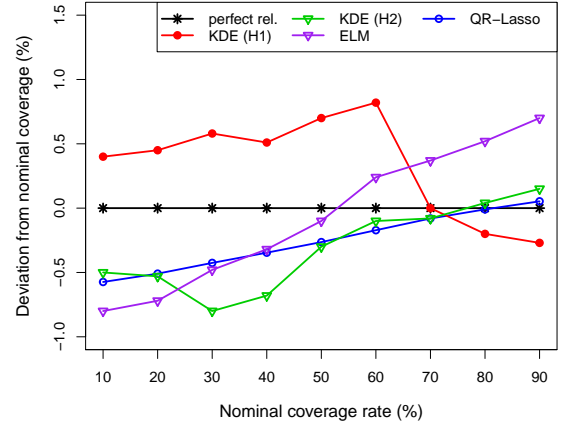


Fig. 6. Reliability of the predictive densities for 3-hour horizons. The red and green lines represent the KDE model with two cases of smoothing matrix. The purple and blue lines represent the ELM and QR-Lasso models. The values are given as average over the forecasting length on the testing set.

zons. Perfect reliability means that there is no deviation, and this is represented by the horizontal black line on the figure. The deviation of the ELM model has also been plotted. Two deviations are plotted for the KDE representing two different smoothing matrices H_1, H_2 . These smoothing matrices are respectively the results of the unbiased cross validation and the smooth cross-validation method (see [56] for more details on the functions and minimization process). The figure shows that, depending on the values of the smoothing matrices, the predictive densities can be either over-estimated (red line for KDE/H1) or under-estimated (green line for KDE/H2). The values of the deviation for the QR-Lasso model do not exceed $\pm 0.5\%$ and are lower than those of the KDE model. Moreover, the deviations of the nominal reliability values of the QR-Lasso model are lower than those of the ELM model (purple line).

Sharpness is a complementary analysis tool to reliability and can be used to evaluate the concentration of the predictive distributions. Sharpness is often evaluated by taking the average width of centered prediction intervals. After defining a set of prediction intervals, the distances between the two boundaries (length of interval) are computed and all of these distance values are averaged over the evaluation set [63]. The more concentrated the predictive distributions are, the sharper the forecasts, which is preferable. The sharpness is defined for a quantile level α for the horizon h as:

$$sharp_h^\alpha = \frac{1}{N} \sum_{t=1}^N \left(\hat{q}_{t+h|t}^{1-\frac{\alpha}{2}} - \hat{q}_{t+h|t}^{\frac{\alpha}{2}} \right). \quad (15)$$

The evaluation of sharpness (as inter-quantile interval lengths) is presented in Figure 7 for the same prediction intervals as on the reliability plot. The maximum observed production is used to normalize the interval lengths. The figure shows that the average interval length increases with the nominal quantile (coverage rate). The values range from 3% to 58% and are lower than those observed for the KDE model for quantiles up to q^{60} . The inter-quantile interval lengths in the QR-Lasso

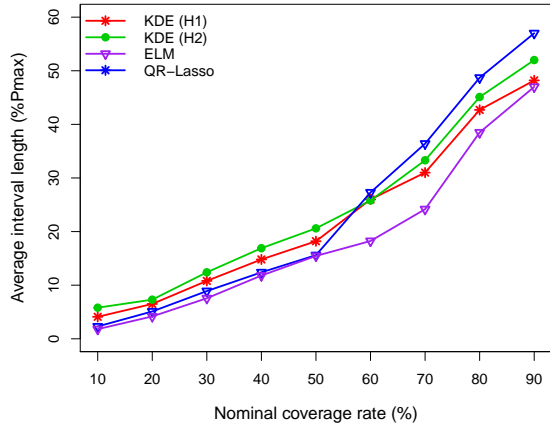


Fig. 7. Sharpness evaluation of the predictive densities for 3-hour horizons. The red and green lines are the interval lengths of the KDE model for two cases of smoothing matrix. The purple and blue lines are those of the ELM and QR-Lasso models.

TABLE II
CRPS OF THE SPATIO-TEMPORAL QR-LASSO MODEL AND THE REFERENCE KDE MODEL FOR FIVE POWER PLANTS.

Power plant	CRPS (% Pmax)		
	KDE	ELM	QR-Lasso
P1	7.20	7.00	5.20
P2	8.32	7.42	4.14
P3	7.65	7.57	5.23
P4	8.67	7.68	5.14
P5	8.32	7.08	4.38

model are higher than those in the KDE model for upper quantiles.

D. Overall Probabilistic Forecasting Skill

In addition to reliability and sharpness, an overall evaluation criterion for probabilistic forecasts is the Continuous Rank Probability Score (CRPS). The CRPS evaluates the entire predictive distribution and can be seen as a criterion that combines reliability with sharpness. It is defined for a cumulative distribution function F and its observation y as

$$crps(F, y) = \int_{-\infty}^{\infty} (F(x) - H(x - y))^2 dx \quad (16)$$

where $H(x)$ is the Heaviside function, whose value is 0 for strictly negative x and 1 for positive x . The CRPS can be interpreted as a measure of distance between the observed cumulative distribution function and the forecasted one [64].

The average CRPS values over the 6-hour forecast horizon are presented in Table II for the reference KDE model with matrix $H1$ (the one with the best performances), the ELM model and the QR-Lasso model for five power plants in the test case. The CRPS values are lowest for the QR-Lasso model, confirming the improved forecasting performance compared to the KDE and ELM models. This analysis provides one more illustration of how well the QR-Lasso model performs. Moreover, the CRPS values are lower than those of the set of models presented in the review [11].

Figure 8 represents the predictive densities as a set of prediction intervals for six days of the testing set for the power plant P_1 . The QR-Lasso model performs quite well to anticipate the variations in production. The length of the quantile intervals is low compared to what can be observed in the literature [47]. Overall, most of the variations for days characterized by high production variability are predicted quite efficiently.

V. CONCLUSION

In this paper we proposed a probabilistic spatio-temporal model that exploits off-site information to improve short-term forecasting of photovoltaic production. The model is based on quantile regression that has been adapted to integrate a LASSO variable selection process. This makes it particularly suitable for situations where data from a high number of PV installations is available since it is able to handle directly the resulting high dimensionality and over-fitting issues that characterize the use of such large amounts of data.

A real world test case with a high number of PV plants was used to illustrate the full potential of the probabilistic spatio-temporal approach. The evaluation of the predicted densities was carried out employing evaluation criteria that are widely used for probabilistic forecasts. The model shows significant performance improvement compared to the reference kernel density model and to models in the literature. The probabilistic spatio-temporal model appears to be an interesting technique that performs well to predict the future densities of PV generation.

The perspectives of this work include the possibility of extending the forecast horizons to daily forecasts. This would emphasize the trade-off significance between NWP forecasts and historic data measurements, and help to define the horizon limit where the spatial-temporal approach does not permit improvement. Another improvement could be to consider a time-varying smoothing matrix to estimate the kernel density, which is a key parameter of this model.

ACKNOWLEDGMENT

The authors would like to thank the French industrial Hespul for providing the PV data as well as the European Center for Medium-Range Weather Forecasts (ECMWF) for providing the NWP data.

REFERENCES

- [1] C. W. Potter, A. Archambault, and K. Westrick, "Building a smarter smart grid through better renewable energy information," in *2009 IEEE/PES Power Systems Conference and Exposition*, Mar. 2009, pp. 1–5.
- [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] G. Kariniotakis, *Renewable Energy Forecasting: From Models to Applications*. Woodhead Publishing, Jun. 2017.
- [4] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, Oct. 2016.
- [5] S. Sobri, S. Koochi-Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," *Energy Conversion and Management*, vol. 156, pp. 459–497, Jan. 2018.

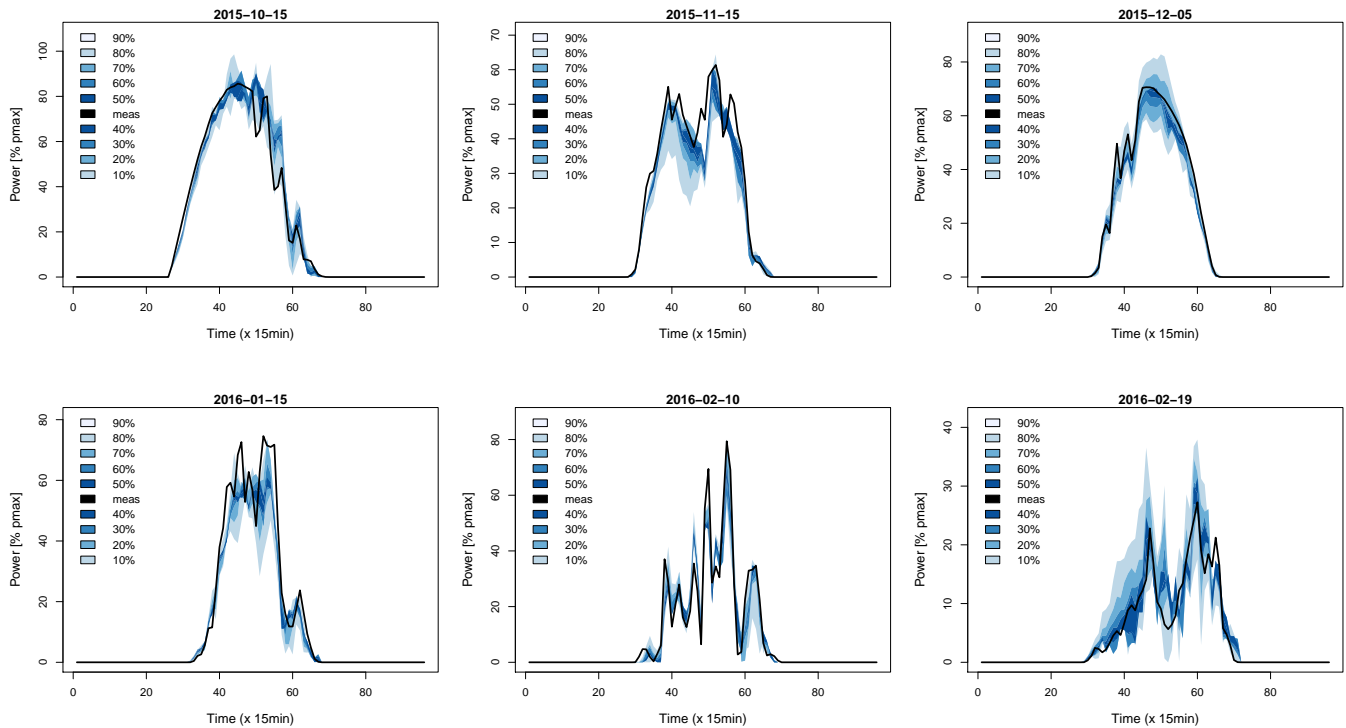


Fig. 8. Examples of predictive densities forecasted with the QR-Lasso model for 6 days in different months of the year. The quantiles are plotted from 10% to 90% with a 10% step. The time step is 15 min and the model is updated every 15 min.

- [6] 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, Jul. 2014.
- [7] S. Sperati, S. Alessandrini, and L. Delle Monache, "An application of the ecmwf ensemble prediction system for short-term solar power forecasting," *Solar Energy*, vol. 133, pp. 437–450, Aug. 2016.
- [8] M. Pierro, F. Bucci, M. De Felice, E. Maggioni, D. Moser, A. Perotto, F. Spada, and C. Cornaro, "Multi-Model Ensemble for day ahead prediction of photovoltaic power generation," *Solar Energy*, vol. 134, pp. 132–146, Sep. 2016.
- [9] Y. Liu, S. Shimada, J. Yoshino, T. Kobayashi, Y. Miwa, and K. Furuta, "Ensemble forecasting of solar irradiance by applying a mesoscale meteorological model," *Solar Energy*, vol. 136, pp. 597–605, Oct. 2016.
- [10] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone, "An analog ensemble for short-term probabilistic solar power forecast," *Applied Energy*, vol. 157, pp. 95–110, Nov. 2015.
- [11] M. Zamo, O. Mestre, P. Arbogast, and O. Pannekoucke, "A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production," *Solar Energy*, vol. 105, pp. 804–816, Jul. 2014.
- [12] G. I. Nagy, G. Barta, S. Kazi, G. Borbély, and G. Simon, "GEFCom2014: Probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1087–1093, Jul. 2016.
- [13] J. Huang and M. Perry, "A semi-empirical approach using gradient boosting and -nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1081–1086, Jul. 2016.
- [14] E. Scolari, F. Sossan, and M. Paolone, "Irradiance prediction intervals for PV stochastic generation in microgrid applications," *Solar Energy*, vol. 139, pp. 116–129, Dec. 2016.
- [15] Y. Chu and C. F. M. Coimbra, "Short-term probabilistic forecasts for Direct Normal Irradiance," *Renewable Energy*, vol. 101, pp. 526–536, Feb. 2017.
- [16] A. Yona, T. Senjyu, A. Y. Saber, T. Funabashi, H. Sekine, and C. H. Kim, "Application of Neural Network to One-Day-Ahead 24 hours Generating Power Forecasting for Photovoltaic System," in *2007 International Conference on Intelligent Systems Applications to Power Systems*, Nov. 2007, pp. 1–6.
- [17] Y. Huang, J. Lu, C. Liu, X. Xu, W. Wang, and X. Zhou, "Comparative study of power forecasting methods for PV stations," in *2010 International Conference on Power System Technology*, Oct. 2010, pp. 1–6.
- [18] A. Mellit and A. 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.
- [19] A. Dolara, F. Grimaccia, S. Leva, M. Mussetta, and E. Ogliari, "A Physical Hybrid Artificial Neural Network for Short Term Forecasting of PV Plant Power Output," *Energies*, vol. 8, no. 2, pp. 1138–1153, Feb. 2015.
- [20] H. Hassani and E. S. Silva, "Forecasting with Big Data: A Review," *Annals of Data Science*, vol. 2, no. 1, pp. 5–19, Mar. 2015.
- [21] S. I. Vagropoulos, E. G. Kardakos, C. K. Simoglou, A. G. Bakirtzis, and J. P. S. Catalão, "ANN-based scenario generation methodology for stochastic variables of electric power systems," *Electric Power Systems Research*, vol. 134, pp. 9–18, May 2016.
- [22] C. Y. Zhang, C. L. P. Chen, M. Gan, and L. Chen, "Predictive Deep Boltzmann Machine for Multiperiod Wind Speed Forecasting," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 4, pp. 1416–1425, Oct. 2015.
- [23] M. Khodayar, O. Kaynak, and M. E. Khodayar, "Rough Deep Neural Architecture for Short-Term Wind Speed Forecasting," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 2770–2779, Dec. 2017.
- [24] 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, Jun. 2008.
- [25] 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.
- [26] R. Amaro e Silva and M. C. Brito, "Impact of network layout and time resolution on spatio-temporal solar forecasting," *Solar Energy*, vol. 163, pp. 329–337, Mar. 2018.
- [27] M. Journée and C. Bertrand, "Improving the spatio-temporal distribution of surface solar radiation data by merging ground and satellite

- measurements," *Remote Sensing of Environment*, vol. 114, no. 11, pp. 2692–2704, Nov. 2010.
- [28] 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.
- [29] S. Quesada-Ruiz, Y. Chu, J. Tovar-Pescador, H. T. C. Pedro, and C. F. M. Coimbra, "Cloud-tracking methodology for intra-hour DNI forecasting," *Solar Energy*, vol. 102, pp. 267–275, Apr. 2014.
- [30] M. Lazzaroni, S. Ferrari, V. Piuri, A. Salman, L. Cristaldi, and M. Faifer, "Models for solar radiation prediction based on different measurement sites," *Measurement*, vol. 63, pp. 346–363, Mar. 2015.
- [31] 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, Sep. 2012.
- [32] E. Lorenz, J. Kühnert, D. Heinemann, K. P. Nielsen, J. Remund, and S. C. Müller, "Comparison of global horizontal irradiance forecasts based on numerical weather prediction models with different spatio-temporal resolutions," *Progress in Photovoltaics: Research and Applications*, vol. 24, no. 12, pp. 1626–1640, Dec. 2016.
- [33] 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.
- [34] F.-V. Gutierrez-Corea, M.-A. Manso-Callejo, M.-P. Moreno-Regidor, and M.-T. Manrique-Sancho, "Forecasting short-term solar irradiance based on artificial neural networks and data from neighboring meteorological stations," *Solar Energy*, vol. 134, pp. 119–131, Sep. 2016.
- [35] Y. He, Q. Xu, J. Wan, and S. Yang, "Short-term power load probability density forecasting based on quantile regression neural network and triangle kernel function," *Energy*, vol. 114, pp. 498–512, Nov. 2016.
- [36] J. Li, J. K. Ward, J. Tong, L. Collins, and G. Platt, "Machine learning for solar irradiance forecasting of photovoltaic system," *Renewable Energy*, vol. 90, pp. 542–553, May 2016.
- [37] 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.
- [38] 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.
- [39] C. Yang and L. Xie, "A novel ARX-based multi-scale spatio-temporal solar power forecast model," in *2012 North American Power Symposium (NAPS)*, Sep. 2012, pp. 1–6.
- [40] V. Berdugo, C. Chaussin, L. Dubus, G. Hebrail, and V. Leboucher, "Analog method for collaborative very-short-term forecasting of power generation from photovoltaic systems," *Next Generation Data Mining Summit (NGDM'11)*, 2011.
- [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.
- [42] F. Golestaneh, H. B. Gooi, and P. Pinson, "Generation and evaluation of space-time trajectories of photovoltaic power," *Applied Energy*, vol. 176, pp. 80–91, Aug. 2016.
- [43] B. Zhang, P. Dehghanian, and M. Kezunovic, "Spatial-temporal solar power forecast through use of Gaussian Conditional Random Fields," in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, Jul. 2016, pp. 1–5.
- [44] X. G. Agoua, R. Girard, and G. Kariniotakis, "Short-term spatio-temporal forecasting of photovoltaic power production," *IEEE Transactions on Sustainable Energy*, vol. PP, no. 99, pp. 1–1, 2017.
- [45] 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.
- [46] A. Tascikaraoglu, B. M. Sanandaji, G. Chicco, V. Cocina, F. Spertino, O. Erdinc, N. G. Paterakis, and J. P. S. Catalão, "A short-term spatio-temporal approach for Photovoltaic power forecasting," in *2016 Power Systems Computation Conference (PSCC)*, Jun. 2016, pp. 1–7.
- [47] F. Golestaneh, P. Pinson, and H. B. Gooi, "Very Short-Term Non-parametric Probabilistic Forecasting of Renewable Energy Generation #x2014; With Application to Solar Energy," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3850–3863, Sep. 2016.
- [48] Y. Zhang and J. Wang, "K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1074–1080, Jul. 2016.
- [49] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Solar Energy*, vol. 83, no. 10, pp. 1772–1783, Oct. 2009.
- [50] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. CRC Press, Apr. 1986.
- [51] G. R. Terrell and D. W. Scott, "Variable Kernel Density Estimation," *The Annals of Statistics*, vol. 20, no. 3, pp. 1236–1265, 1992.
- [52] M. P. Wand and M. C. Jones, *Kernel Smoothing*. CRC Press, Dec. 1994.
- [53] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, Mar. 2015.
- [54] A. R. Mugdadi and I. A. Ahmad, "A bandwidth selection for kernel density estimation of functions of random variables," *Computational Statistics & Data Analysis*, vol. 47, no. 1, pp. 49–62, Aug. 2004.
- [55] T. A. O'Brien, K. Kashinath, N. R. Cavanaugh, W. D. Collins, and J. P. O'Brien, "A fast and objective multidimensional kernel density estimation method: fastKDE," *Computational Statistics & Data Analysis*, vol. 101, pp. 148–160, Sep. 2016.
- [56] T. Duong and M. L. Hazelton, "Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation," *Scandinavian Journal of Statistics*, vol. 32, no. 3, pp. 485–506, Sep. 2005.
- [57] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley, Aug. 1991.
- [58] R. Koehler, *Quantile Regression*. Cambridge University Press, 2005.
- [59] G. B. Dantzig, *Linear Programming and Extensions*, 2007.
- [60] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.
- [61] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2014.
- [62] P. Pinson, H. A. Nielsen, J. K. Møller, H. Madsen, and G. N. Kariniotakis, "Non-parametric probabilistic forecasts of wind power: required properties and evaluation," *Wind Energy*, vol. 10, no. 6, pp. 497–516, Nov. 2007.
- [63] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, Mar. 2007.
- [64] H. Hersbach, "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems," *Weather and Forecasting*, vol. 15, no. 5, pp. 559–570, Oct. 2000.

Xwégnon Ghislain Agoua received a Master's degree in Statistics (2014) from ENSAI (Ecole nationale de la statistique et de l'analyse de l'information), France and a PhD degree (2017) in energetics from MINES ParisTech, PSL - Research University. His research interests include statistical modeling, forecasting techniques, time series analysis, spatio-temporal regression models, and their applications to solar and wind generation modeling.

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

George Kariniotakis (S'95-M'02-SM'11) was born in Athens, Greece. He received his Eng. and M.Sc. degrees from Greece in 1990 and 1992 respectively, and his PhD 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 220 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 time series forecasting, decision-making under uncertainty, modelling, management and power systems planning.