On the simulation of aggregated solar PV forecast errors

Nuño Martinez, Edgar; Koivisto, Matti Juhani; Cutululis, Nicolaos Antonio; Sorensen, Poul

Published in:
I E E E Transactions on Sustainable Energy

Link to article, DOI:
10.1109/TSTE.2018.2818727

Publication date:
2018

Document Version
Peer reviewed version

On the simulation of aggregated solar PV forecast errors

Edgar Nuño*, *Member, IEEE*, Matti Koivisto, Nicolaos Cutululis, *Member, IEEE*, and Poul Sørensen, *Senior Member, IEEE*

***Abstract*—The uncertainty arising from high levels of solar photovoltaic (PV) penetration can have a substantial impact on power system operation. Therefore, there is a need to develop models capable of representing PV generation in a rigorous manner. This paper introduces a novel transformation–based methodology to generate stochastic solar area power forecast scenarios; easy to apply to new locations. We present a simulation study comparing day–ahead solar forecast errors covering regions with different geographical sizes, total installed capacities and climatic characteristics. The results show that our model can capture the spatio–temporal properties and match the long–term statistical properties of actual data. Hence, it can be used to characterize the PV input uncertainty in power system studies.**

***Index Terms*—Solar power generation, forecast uncertainty, simulation, power systems.**

NOMENCLATURE

$\mathcal{A}$	Set of power system areas.
$\mathcal{C}$	Set of area power centres (latitude, longitude).
$\mathcal{D}$	Set of area diameters (km).
$\mathcal{L}$	Set of PV locations.
$\mathcal{S}$	Set of combinations.
Ω	Set of combination weights.
$\boldsymbol{x_t}$	Vector of PV power measurements time series (MW).
$\widetilde{\boldsymbol{x}}_{\boldsymbol{t}}$	Vector of PV power forecasts time series (MW).
$\widetilde{\boldsymbol{x}}_{\boldsymbol{t}}^{sim}$	Vector of simulated PV power forecast time series scenarios (MW).
$\boldsymbol{\tau_t}$	Vector of clear–sky power time series (MW).
$\boldsymbol{\xi_t}$	Vector of transformed PV forecast error time series (p.u.).
f^{h}_{autocorr}	Function describing the evolution of the hth forecast error autocorrelation lag with the area size.
f_{corr}	Functional relationship between the forecast error correlation coefficient with the distance between areas.
DNI	Direct Normal Irradiance (W/m^2).
DHI	Diffuse Horizontal Irradiance (W/m^2).

I. INTRODUCTION

LIMITED predictability and variability are two fundamental characteristics associated to photovoltaic (PV) solar power. They generally translate into errors in the estimated production, which can represent a challenge during the operation of highly solar–penetrated power systems. Solar forecasting is a fast–growing field, however, the non-linear and stochastic effect of cloud motion on solar irradiation limits the definition of accurate and robust all-purpose forecasting systems [1]. In this regard, different forecasting horizons e.g. from minutes to 6 hours and up to a few days ahead require dedicated techniques [2]. Among all the available commercial products, day-ahead forecasts represent an important operational planning tool in power systems. They are generally based on numerical weather prediction (NWP) models [2] and are used as an input during the market-clearing process in order to optimize the generator's scheduling for the next day of operation on an hourly resolution; matching the current electricity market designs in Europe. Errors in day–ahead forecasts will need to be corrected during the different intra–day markets to prevent eventual imbalances between generation and demand. However, this situation can translate into a sub–optimal result mainly due to the limited liquidity of such markets.

Power system operators are normally interested in aggregated forecasts at a regional level. Generally, wind and solar power forecasts are performed for a group of representative installations and later up-scaled to represent the aggregated production in the system [3, 4]. The uncertainty associated to operational forecasts increases with the leading time due to the complex nature of the atmosphere. A natural way to minimize this shortcoming is to move away from point estimations into probabilistic scenarios. For instance, they can be generated by slightly changing the parametrization or starting points of NWP models [5] or via chains of conditional distributions [6, 7, 8, 9, 10, 11]. The first technique, known as ensemble forecasting, is based on deterministic equations and generally comes at the expense of large computation times. The latter requires a significant amount of information, namely the conditional probability distribution of each prediction step for every location. Some power system studies e.g. integration, market studies, etc. may not require the best possible prediction, but realistic forecast errors to study the impact of renewable energy sources (RES) on power systems. Often, these errors are assumed to be constant as a percentage of the installed capacity or to show an increasing standard deviation changing with the leading time [12]. Söder [13] firstly addressed this issue by developing a methodology to simulate wind speed forecast errors at different areas of the power system using a Vector Autoregressive Moving Average (1,1) (VARMA) model. The method has been used

The authors are with the Department of Wind Energy, Technical University of Denmark, Denmark (*corresponding author e-mail: ednu@dtu.dk).

for different applications such as stochastic optimization [14]. Different methods based on time series have been proposed to forecast PV generation [15], often relying on observations from nearby locations[16, 17]. However, to the best of our knowledge the simulation of stochastic PV forecast scenarios as such has not yet received the same degree of attention.

The objective of this paper is twofold. First, to generalize the methodology proposed in [18] to simulate statistical day-ahead aggregated PV forecast scenarios and second, to validate the results considering areas with different weather characteristics and levels of aggregation. The paper is organized as follows. Section II presents the general methodology and the theoretical basis of this work. A case study along with the model calibration results are summarized in section III. Furthermore, validation results are presented and discussed in sections IV and V respectively, followed by some concluding remarks in section VI.

II. Methodology

Solar generation is mainly determined by the availability of the solar resource as well as cloud patterns. Moreover, it can be significantly affected by additional local phenomena such as suspended solids particles and shading effects. These conditions hinder the ability to formulate a valid model able to perform well at locations with different climatic characteristics, since the performance of the same module can significantly differ [19]. In this work, we describe a stochastic model able to reproduce day-ahead forecast scenarios understood in a statistical sense i.e. focusing on the statistics of the forecast errors. Thus, the main goal is to simulate realistic realizations matching the most important properties of the actual forecast errors e.g. their autocorrelation, cross-correlation and distributional characteristics, rather than accurate predictions. The fundamental motivation behind this work lies in the fact that stochastic methods offer much higher computational efficiency than traditional NWP tools. For instance, it is possible to simulate predictions over a year for a given location within seconds on a regular laptop computer; without the need to rely on high performance computer (HPC) clusters. Thus, developing a method to accurately reproduce the main statistical features of state-of-the-art predictions at a limited computational cost can foster the application of probabilistic techniques in power systems.

As proposed in [20, 21], the time series of interest can be firstly transformed into a stationary multivariate Gaussian process to capture both temporal and spatial dependencies in the Gaussian domain. This procedure allows for a straightforward simulation of data, which can then be transformed back to the domain on interest. The proposed methodology is summarized in Fig. 1. Input and output variables at each stage of the process are explained in the following sections.

A. Geographical description

Photovoltaic installed capacity at the regional level is generally distributed among a large number of small generation facilities. As a matter of fact, the spatial distribution of the panels can significantly affect the properties of the aggregated production and prediction, also known as smoothing effect [22]. The geographical properties of each individual area $A_i \in \mathcal{A} = \{A_1, \ldots, A_k\}$ can be summarized using a single quantity, from here on referred as the *area diameter*, i.e. $D_i \in \mathcal{D} = \{D_1, \ldots, D_k\}$. It can be defined as the average distance (in kilometres) between all the installations in that area $L_i \subset \mathcal{L}$ and the geographical point minimizing the distance between them; known as its *power centre*, i.e. $C_i \in \mathcal{C}$. Note that all distances calculated from individual locations are always weighted by its rated power. In a nutshell, high diameter values indicate that the panels are relatively spread; hence, as an aggregate, less exposed to local weather phenomena. Conversely, low diameters reveal that PV panels are geographically closer to each other within the area.

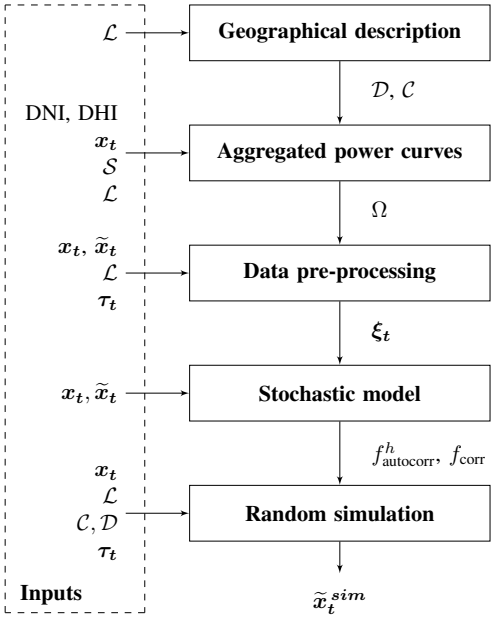

Fig. 1: Schematic diagram of the scenario simulation process

B. Aggregated power curves

One of the biggest challenges of large–scale PV modelling correspond to the limited available information regarding the location, inclination and orientation of the individual PV panels. Inspired on current wind power modelling techniques [23], we applied a method based on aggregated solar power curves [18, 24] to account for the spatial distribution, orientation and inclination uncertainty. We consider a set of combinations $\mathcal{S} = \{S_1, \ldots, S_n\}$ corresponding to different azimuth and tilt angle pairs for all the installations inside each area. Every combination has a specific weight associated, so for a given area i the set $\Omega_i = \{\omega_{1,i}, \ldots, \omega_{n,i}\}$ includes all the combination weights. Note that these sets of weights might change across areas i.e. $\Omega_1 \neq \Omega_2 \neq \ldots \neq \Omega_k$. Also, let $\Omega = \{\Omega_1, \ldots, \Omega_k\}$ be the set of weights across all the areas of interest.

Let now $x_{i,t} = [x_{i,1}, \dots x_{i,T}]$ be the time series of measured aggregated solar production at area i of the system. The production at each area is directly related to the average global effective irradiance reaching its surface g. It can be computed as a function of the selected combinations $\mathcal{S}$ as well as the meteorological information near all the reported locations e.g. direct normal irradiance (DNI) and diffuse horizontal irradiance (DHI); obtained for example using NWP–based reanalysis techniques [25]. For a given combination j and a specific area i, let us define an *aggregated solar power curve* [26] as a function in $\mathbb{R}$ mapping time series of the area–averaged effective global irradiance corresponding to that combination and the measured solar production time series:

$$\tilde{\gamma}_{j,i} : \ g_{j,i,t} \rightarrow x_{i,t} \tag{1}$$

Based on the empirical data analysed, a linear relationship was found satisfactory to describe this relationship. Consequently, if we operate backwards and apply the previously derived power curve to a time series of effective irradiance, it is possible to obtain the corresponding estimated production for combination j and area i:

$$x'_{j,i,t} = \tilde{\gamma}_{j,i}(g_{j,i,t}) \tag{2}$$

Let us now assume that the true aggregated power production at area i can be approximated as:

$$x_{i,t} \simeq \sum_{j=1}^{n} \omega_{j,i} \cdot x'_{j,i,t} \tag{3}$$

Then, the weights of the possible inclination and orientation specifications for that area, $\Omega_i = \{\omega_{1,i}, \dots, \omega_{n,i}\}$, can be estimated by fitting a multiple regression model, where the independent variables are the estimated PV generation time series of the combinations $x'_{i,j,t}$ and the dependent variable is the measured (known) PV generation of the area. The weights are constrained to be positive, as generation from any of the inclination and orientation combinations can only be positive. The magnitude of each individual weight is related to the unknown orientation and inclination of all the PV installations mix inside an area. The higher the weight, the closer the theoretical production of that area with all the installations with the azimuth and tilt angles corresponding to that combination would be.

C. Data pre-processing

Solar generation depends on the solar irradiance at the ground level, which at hourly resolution will show both strong daily and yearly patterns due to the the Earth's rotation and translation respectively. This deterministic profile translates into a changing variance of the solar forecast errors across the day. We propose to correct this highly heteroscedastic random process by pre-processing the raw data into the standard normal space by including several transformations. Let $\boldsymbol{x_t} = [x_{1,t}, \dots, x_{k,t}]^{\mathrm{T}}$ be the random vector of solar measurements time series at areas $i = 1, \dots, k$, and let $\widetilde{\boldsymbol{x}}_{\boldsymbol{t}} = [\tilde{x}_{1,t}, \dots \widetilde{x}_{k,t}]^{\mathrm{T}}$ be the vector of solar PV forecast time series. Both solar vectors must be divided by the maximum possible generation time series under clear–sky conditions [27], hereafter referred as clear–sky power $\boldsymbol{\tau_t} = [\tau_{1,t}, \dots, \tau_{k,t}]^{\mathrm{T}}$, so the stochastic component can be isolated from the original signal. Note that the clear–sky power is specified here so that it gets values bounded between 0 and 1. The normalized time series for any $i = 1, \dots, k$ are:

$$\begin{aligned} m_{i,t} &= \frac{y_{i,t}}{\tau_{i,t}} \in [0,1] \\ \widetilde{m}_{i,t} &= \frac{\widetilde{y}_{i,t}}{\tau_{i,t}} \in [0,1] \end{aligned} \tag{4}$$

where $y_{i,t}$ and $\widetilde{y}_{i,t}$ correspond to the time series of PV measurements and PV forecasts at area i, normalized by the hourly installed capacity in the region e.g. $y_{i,t} = x_{i,t}/p_{i,t}$. The last step in the pre-processing stage requires transforming the previously defined variables into new normally distributed random variables:

$$\begin{aligned} z_{i,t} &= \Phi^{-1}\big[F_i(m_{i,t})\big] \sim \mathcal{N}(0,1) \\ \tilde{z}_{i,t} &= \Phi^{-1}\big[\widetilde{F}_i(\widetilde{m}_{i,t})\big] \sim \mathcal{N}(0,1) \end{aligned} \tag{5}$$

where Φ^{-1} is the inverse cumulative distribution function (CDF) of a standard normal distribution and F_i, $\widetilde{F}_i$ represent the empirical area CDFs of the corresponding normalized variables $m_{i,t}$ and $\widetilde{m}_{i,t}$ respectively. Consequently, the vector of forecast errors can be defined as:

$$\boldsymbol{\xi_t} = \widetilde{\boldsymbol{z}}_{\boldsymbol{t}} - \boldsymbol{z_t} \ \in \mathbb{R}^{\mathrm{k}} \tag{6}$$

D. Stochastic model

Multivariate Vector Autoregressive Moving Average (VARMA) models represent a robust tool to account for both the temporal and spatial dependencies between random variables, such is the case for solar forecast errors. More specifically, we found vector autoregressive (VAR) processes as good candidates to capture the variability of the solar PV forecast errors at the power system area level based on initial investigations [18]. They can be estimated based on ordinary least squares (OLS); which significantly reduces the complexity of their estimation compared to other models relying on maximum likelihood estimation (MLE) e.g. when a moving average term is added. In a general way, the evolution of a random vector such in (6), $\boldsymbol{\xi_t} = [\xi_{1,t}, \dots, \xi_{k,t}]^{\mathrm{T}}$ at time t following a H–th order VAR(H) process can be presented as:

$$\boldsymbol{\xi_t} = \boldsymbol{\alpha} + \sum_{h=1}^{H} \mathbf{B^h} \cdot \boldsymbol{\xi_{t-h}} + \boldsymbol{\epsilon}_t \tag{7}$$

where $\boldsymbol{\alpha}$ is a vector of additive constants, $\boldsymbol{\xi_{t-h}}$ is a vector of past realizations of $\boldsymbol{\xi_t}$ at lag h, $\mathbf{B^h}$ is a matrix of coefficients corresponding to that lag and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is the random innovation term with zero mean and correlation matrix $\boldsymbol{\Sigma}$. Both temporal and spatial properties are related to the underlying multivariate time series process. For simplicity, we assumed a diagonal structure of the coefficient matrices i.e.:

$$\mathbf{B^h} = \begin{bmatrix} b_1^h & & 0 \\ & \ddots & \\ 0 & & b_k^h \end{bmatrix} \tag{8}$$

An optimization routine was proposed in [13] in order to construct $\mathbf{B}^{\boldsymbol{h}}$ so that the root mean square error (RMSE) of the simulated forecast errors was minimized. This approach was relatively limited since it assumed a single set of parameters i.e. $b_1^h = b_2^h = \ldots = b_k^h$ and did not directly aimed at capturing the autocorrelation structure of the data. Alternatively, we explicitly connect the temporal properties of the aggregated forecast errors to the spatial characteristics of the ith area, summarized by its diameter $\mathcal{D}_i$. Consequently, knowing the autocorrelation coefficients for each different lag i.e. $r_i^1, \ldots, r_i^h$, it is possible to operate backwards and estimate the elements of the coefficient matrices of the model corresponding to each area via the Yule–Walker eqs. [28].

$$\begin{bmatrix} b_i^1 \\ b_i^2 \\ \vdots \\ b_i^{h-1} \\ b_i^h \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ r_i^1 & 1 & & \text{Sym.} & \\ \vdots & \vdots & \ddots & & \\ r_i^{h-2} & r_i^{h-3} & \ldots & 1 & \\ r_i^{h-1} & r_i^{h-2} & \ldots & r_i^1 & 1 \end{bmatrix}^{-1} \times \begin{bmatrix} r_i^1 \\ r_i^2 \\ \vdots \\ r_i^{h-1} \\ r_i^h \end{bmatrix} \quad (9)$$

where the H–lag autocorrelation coefficients depend on the area characteristics:

$$r_i^h = f_{\text{autocorr}}^h(D_i) \quad (10)$$

Hence, repeating the process over for all the areas $\{A_1, \ldots, A_k\} \in \mathcal{A}$, it is possible to build a set of model parameter matrices corresponding to each individual lag $\boldsymbol{B}^{\boldsymbol{h}}$ for $h = 1, \ldots, H$, as in eq. (8). The geographical dependence structure between forecast errors of the different areas will be determined by the correlation matrix of the innovation term, $\boldsymbol{\Sigma}$. Based on the existing literature [29, 30], we assumed a monotonically decreasing relationship between the pairwise zero-lag cross-correlation (also known as Pearson's correlation coefficient) between the forecast errors at two different areas and their geographical distance.

$$\rho_{i,j} = f_{\text{corr}}(d(C_i, C_j)) \quad (11)$$

where C_i, C_j correspond to the respective power centres of area i and j. The estimated shape of f_{corr} is presented in section III-C. Note that eq. (11) does not guarantee a suitable correlation matrix, specially as dimension increases. In such cases, there are different methods to obtain a positive-definite correlation matrix as close as possible to the empirical one e.g. [31].

E. Random simulation

Once the individual $f_{\text{autocorr}}^h(\mathcal{D})$ function values associated with each h lag and the relationship between cross-correlation and distance i.e. $f_{\text{corr}}(\mathcal{C})$ have been defined, the random vector of forecast errors $\boldsymbol{\xi}_{\boldsymbol{t}}^{\boldsymbol{sim}} = [\xi_{1,t}^{sim}, \ldots, \xi_{k,t}^{sim}]^{\mathrm{T}}$ can be simulated according to (7). Subsequently, it has to be transformed back to the original domain. First, the simulated forecast vector can be calculated based on (6):

$$\widetilde{\boldsymbol{z}}_{\boldsymbol{t}}^{\boldsymbol{sim}} = \boldsymbol{\xi}_{\boldsymbol{t}}^{\boldsymbol{sim}} + \boldsymbol{z_t} \quad \in \mathbb{R}^{\mathrm{k}} \quad (12)$$

and then the inverse of eqs. (4), (5) can be applied:

$$\begin{aligned} \widetilde{y}_{i,t}^{sim} &= F_i^{-1}\big[\Phi(\widetilde{z}_{i,t}^{sim})\big] \cdot \tau_{i,t} \\ \widetilde{x}_{i,t}^{sim} &= \widetilde{y}_{i,t}^{sim} \cdot p_{i,t} \end{aligned} \quad (13)$$

Note that $\widetilde{\boldsymbol{x}}_{\boldsymbol{t}}^{\boldsymbol{sim}} = [\widetilde{x}_{1,t}^{sim}, \ldots, \widetilde{x}_{i,t}^{sim}]$ is the vector including the forecast scenarios for all the areas in the system. In this case, we consider that the empirical distribution of the forecasts can be approximated by the distribution of the measured counterparts. Therefore, no prior forecast data are required to simulate any forecast scenario. The overall process can be summarized as follows:

- *Step* 1 – Characterize each individual area calculating their power centres $\mathcal{C}$ and diameters $\mathcal{D}$.
- *Step* 2 – Determine the scenario weights Ω matching the measured power against the effective irradiance for each of the inclination and orientation combinations.
- *Step* 3 – Pre–process the original data (4) – (5) and calculate the vector of transformed forecast errors $\boldsymbol{\xi_t}$, (6).
- *Step* 4 – Derive the functions f_{autocorr}^h for $h = 1, \ldots, \mathrm{H}$ lags included in the stochastic VAR(H) model, as well as f_{corr}.
- *Step* 5 – Based on $\mathcal{D}$ and f_{autocorr}^h obtained in *Steps 1,4* estimate the lag–h autocorrelation coefficients and derive the model parameter matrices applying (9).
- *Step* 6 – Construct a cross-correlation matrix considering the distances between locations (11).
- *Step* 7 – Simulate the VAR(H) random process.
- *Step* 8 – Apply (12) – (13) to obtain the final forecast scenario.

III. Case Study

A. PV power areas

We selected two different climate regions in order to test the performance of the proposed model: one corresponding to a relatively sunny and stable conditions (Southern Europe) and another one presenting a poorer solar irradiation resource (Central Europe). Table I summarizes the characteristics of each of the studied regions and Fig. 2 shows an overview of the spatial distribution of the PV panels in the areas. From the initial pool of locations, we selected 17 different regions for calibrating the model parameters, leaving 6 for validation. We gathered publicly available solar production data and day–ahead forecasts for each region from the individual TSOs: *Elia* [32], *Amprion* [33], *TenneT* [34] and *Transnet BW* [35], as well as the *European Network of Transmission System Operators for Electricity* (ENTSO-E) [36]. Additionally, we inferred the approximate geographical distribution of the PV panels per postal code based on the information from the national energy agencies: *Bundesnetzagentur* (BNetzA) [37], *Gestore dei Servizi Energetici* (GSE) [38], *Commission wallonne pour l'Energie* (CWaPE) [39] and *Vlaamse Regulator van de Elektriciteits- en Gasmarkt* (VREG) [40].

B. Meteorological data

We run the Weather Research and Forecasting (WRF) [41] model, a widely used open-source mesoscale modelling system to simulate the meteorological conditions matching the historical power records in the areas. We used time series of DNI and DHI on a 10 by 10 kilometre spatial grid and a temporal resolution of one hour. An analogous method for generating

Fig. 2: Geographical distribution of the PV panels in Southern (left) and central Europe (right). Individual plants are represented by grey dots. The power center of the areas used in calibration are displayed in blue and the validation areas in red.

TABLE I: Regional PV characteristics

Region	Country	D (km)	Capacity MW (end 2015)
Antwerp	BE	21.3	1,981
Brussels		<10	194
East Flanders		21.2	1,924
Flemish Brabant		22.1	1,013
Hainaut		35.0	765
Liege		17.0	1,078
Limburg		20.3	1,840
Luxembourg*		32.2	344
Namur		21.1	468
Walloon Brabant		13.6	289
West Flanders*		27.7	1,913
Bayern*	GE	101.1	11,309
Schleswig-Holstein*		55.2	1,498
Hessen		61.9	1,811
Bremen, L. Saxony		98.8	3,622
Amprion		101.8	6,700
Baden-Württemberg		71.6	5,117
Centro-Sud	IT	101.9	2,654
Sud		104.7	3,613
Centro-Nord*		82.0	2,271
Nord		130.0	8,319
Sicily		77.3	1,309
Sardinia*		63.0	726

$*$ Used for validation

time series was used and verified in [25]. We derived the global horizontal irradiance (GHI) from the previous variables and applied the Haurwitz model [42] to derive the GHI under clear–sky conditions GHI^{clear} considering the solar geometry at the middle of each hour. Lastly, under each combination j at area i, we used the Perez model [43] and trigonometric relations to calculate the effective and clear–sky effective irradiance time series i.e. $g_{j,i,t}$ and $g^{clear}_{j,i,t}$ respectively. Each time series of effective irradiance was used to derive empirically the $k \times n$ aggregated power curves as in (1). These power curves were later applied to obtain the set of combination weights Ω via eqs. (2) – (3). Subsequently, the average clear–sky effective irradiance for area $i = 1, \ldots, k$ was calculated as:

$$g^{clear}_{i,t} = \sum_{j=1}^{n} \omega_{j,i} \cdot g^{clear}_{j,i,t} \tag{14}$$

and the conversion to clear–sky power was performed assuming an ideal power curve i.e. the rated power of the area corresponds to 1,000 W/m^2:

$$\tau_{i,t} = g^{clear}_{i,t} / 1,000 \tag{15}$$

C. Model calibration

We initially selected a VAR(2) model aiming at reducing the complexity of the random vector simulation model. Fig. 3 presents the first and second–lag autocorrelation coefficients of the transformed area forecast errors ξ, as well as those corresponding to different combinations of the original areas based on geographical proximity.

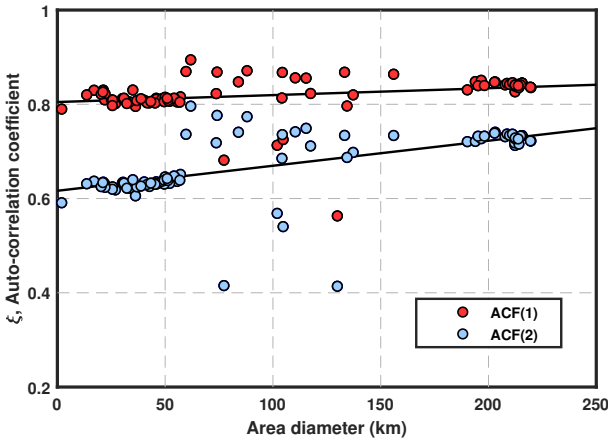

Fig. 3: Evolution of the lag–1, ACF (1) and lag–2, ACF (2), autocorrelation coefficients with the area diameter parameter.

The lag–1 coefficient clearly remains relatively stable as the area diameter increases. However, the lag–2 coefficient tends to increase with the diameter due to the smoothing effect caused by the geographical spread of the installations. Therefore, we chose to approximate eqs. (10) as linear functions. Similarly, Fig. 4 illustrates how the standard deviation of the transformed forecast errors changes with the diameter of the area.

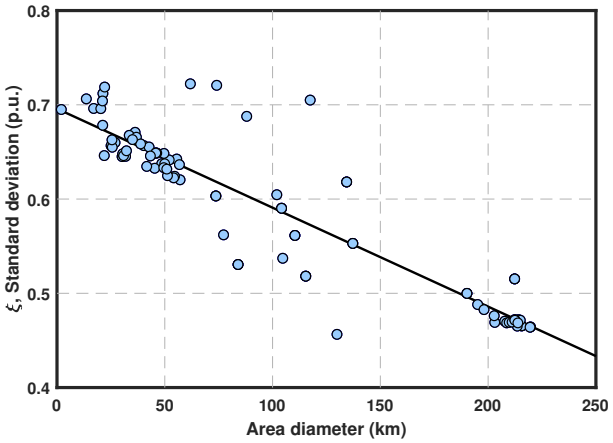

Fig. 4: Change in the standard deviation of the forecast errors ξ as the areas parameter increases.

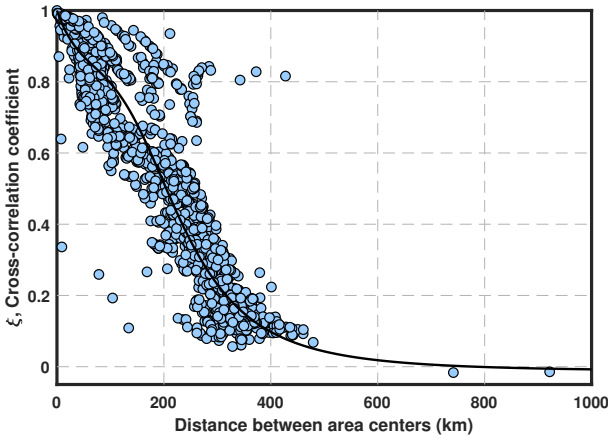

Fig. 5: Pairwise cross-correlation coefficient between area transformed forecast errors ξ as a function of the distance between areas.

The standard deviation of ξ significantly decreases as the PV installations are more spread across the area. Similarly to Fig. 3, we found two well–defined clusters. One between 0 and 60 kilometres corresponding to the Belgian areas and their combinations and a second cluster near 200 kilometres formed by the combination of the areas in Germany and Belgium. The individual areas were aggregated based on their geographical proximity. Hence, three out of the four Italian areas used in the calibration of the model were combined separately. Finally, Fig. 5 shows the relationship between the cross–correlation of the transformed forecast errors and the inter–area geographical distance. The results suggest that the cross-correlation between the transformed forecast errors decreases as the distance between the area centres increases. Moreover, the cross correlation tends to be slightly negative for distances greater than 700 kilometres. The relationship was clearly not linear. Instead, we found that a third–order rational function (shown in black) was able to match the cross–correlation of the transformed forecast errors quite accurately, i.e.:

$$\rho_{i,j} = \frac{\delta_1}{\delta_2 + d_{i,j} + \delta_3 \cdot d_{i,j}^2 + \delta_4 \cdot d_{i,j}^3} \tag{16}$$

where δ_1, δ_2, δ_3 and δ_4 are constants and $d_{i,j} = d(C_i, C_j)$ represents the Euclidean distance between the power centres of area i and area j.

IV. Results

This section presents the results from the six regions included in the validation set, as shown in Fig. 2. For each case, we used information regarding the spatial distribution of the panels across the areas $\mathcal{D}, \mathcal{C}, \mathcal{L}$, the clear–sky power vector $\boldsymbol{\tau_t}$ based on set of combination weights allocated to each area as well as the actual measurements vector time series $\boldsymbol{x_t}$ to simulate day-ahead forecast scenarios for a full year on an hourly resolution in a single run. The CPU time required to simulate 500 realizations on a two–core

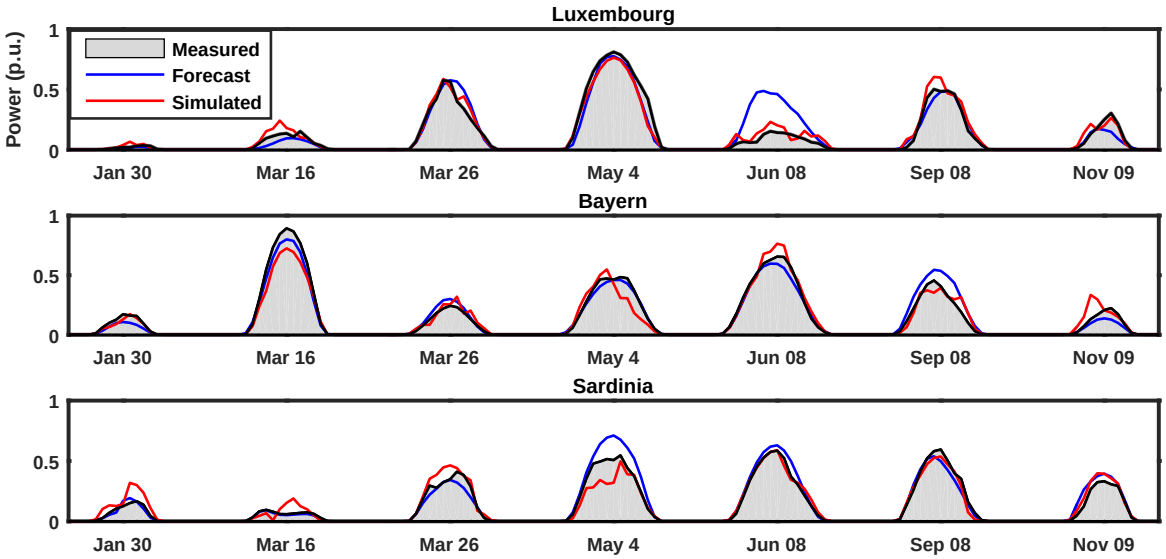

Fig. 6: Comparison of the measured PV production (grey area), actual forecast (blue) and one model run (red) for selected days of the simulated year.

computer with 2.10 GHz and 2.70 GHz and 8 GB RAM was 132.4 seconds. Power measurements (grey shaded area) and actual forecasts (blue lines) for different days of the year are presented in Fig. 6. Moreover, the results of a single model run are presented in red. Simulated forecasts were relatively smooth, resembling characteristics of aggregated production. In addition, our model was able to capture yearly trends observed in the measurements as well as the variation in the sunshine duration throughout the year. Furthermore, Fig. 7

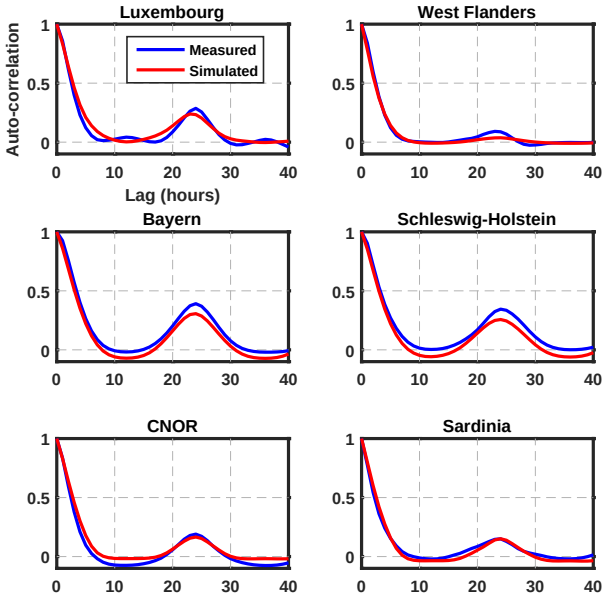

Fig. 7: Auto–correlation function of the forecast errors of the regions included in the validation set. Measurements and simulated scenarios are represented by blue and red lines respectively.

illustrates the auto–correlation function of the actual (blue) and the mean of the simulated (red) forecast errors based on 500 model realizations. It can be observed that, despite its simplicity, the proposed VAR(2) model was able to closely match the temporal properties of the forecast errors. This highlights the importance of the pre-processing steps (5) in order to successfully isolate the stochastic component of the original measurement and forecast vectors. We also observed an intra–day temporal structure, exemplified by a 24–lag peak in the forecast error autocorrelation function. Even though this peak was significant, specially in Luxembourg, Bayern and Hessen, it was captured during the back–transformation step (12). The same conversion also ensured that the forecast never exceeds the theoretical maximum production, defined by the clear–sky power or goes below zero. According to (13), for each area $\widetilde{m}_{i,t}^{sim} \in [0,1]$, hence $\widetilde{y}_{i,t}^{sim} \in [0, \max(\tau_{i,t})]$. In addition, Fig. 8 shows the evolution of the zero–lag cross-correlation coefficient of the measured forecast errors and a 95% confidence band for the simulated forecast errors based on 500 realizations. It can be observed that most of the measurements laid within the confidence band. The representation of the geographical structure of the model can be further improved by including off–diagonal terms in the parameter matrices (8). This will require to consider individual cross–correlation functions for the different lags of the model, similar to the selected procedure applied for the autocorrelation coefficients. We were also interested in reproducing the

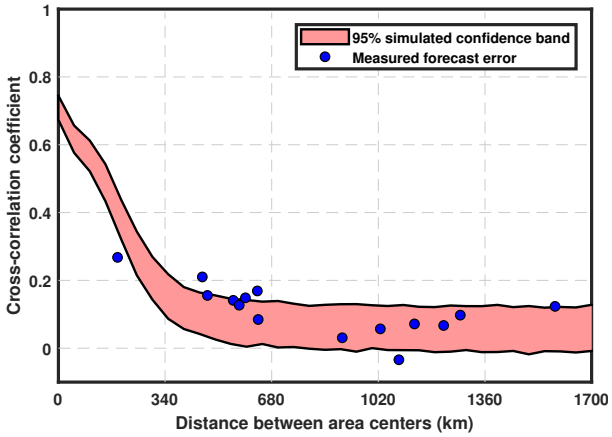

Fig. 8: Cross-correlation of the forecast errors as a function of the distance between area centres. The blue points correspond to the measured cross-correlation values whereas the red area represents the 95% confidence band of the simulations.

distributional properties of the actual forecast errors. Fig. 9 illustrates the histogram of the measured (blue) and simulated (red) forecast errors for the validation set using 0.05 per unit bins. The proposed methodology does not directly preserve

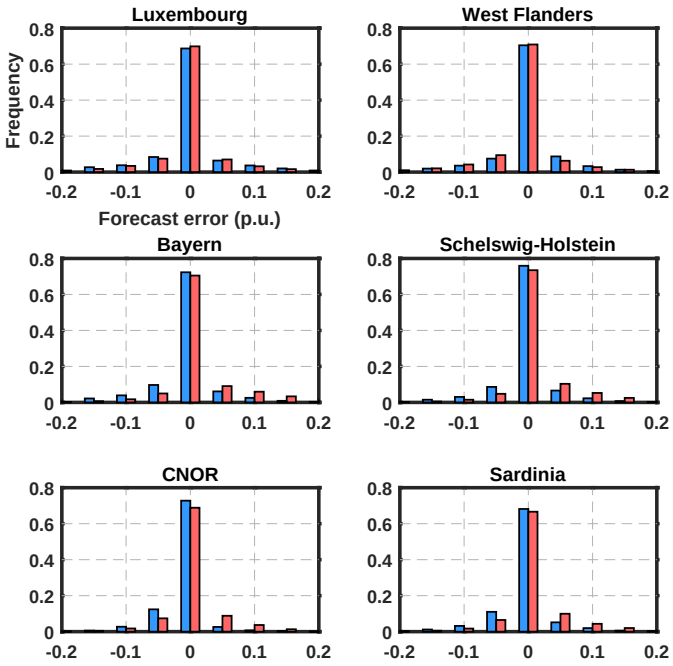

Fig. 9: Histograms of the measured (blue) and simulated (red) area forecast errors.

the statistical distribution of the forecast errors, but rather the distribution of the PV forecast scenarios. Nevertheless, we observed that the simulated results precisely matched the actual data, supporting the main assumption in (13) i.e. $F_i \simeq \widetilde{F}_i$. In order to assess the accuracy of the simulations, Table II summarizes the mean standard deviation σ and the

TABLE II: Summary statistics of the simulated and measured forecast errors

Region	σ	nRMSE	CRPS	$Q_{2.5}$	$Q_{97.5}$
Luxembourg	0.099 (0.068)	0.122 (0.084)	0.0225	-0.27 (-0.14)	0.15 (0.19)
West Flanders	0.059 (0.061)	0.073 (0.075)	0.0093	-0.12 (-0.12)	0.16 (0.16)
Bayern	0.066 (0.054)	0.073 (0.058)	0.0164	-0.20 (-0.10)	0.08 (0.15)
Schleswig-Holstein	0.052 (0.045)	0.081 (0.068)	0.0131	-0.15 (-0.09)	0.07 (0.12)
Centro-Nord	0.051 (0.031)	0.082 (0.051)	0.0093	-0.12 (-0.04)	0.08 (0.09)
Sardinia	0.057 (0.044)	0.079 (0.061)	0.0122	-0.15 (-0.09)	0.08 (0.11)

All the results are expressed in per unit. The parenthesis values correspond to the actual measured forecast errors.

normalized RMSE (nRMSE) using the range of the data at each location as normalization factor. Note that the results are based on 500 forecast scenario realizations. Furthermore, the continuous rank probability score (CRPS) [44] was added as a metric evaluating the simulated forecast error density. The quantities between parenthesis correspond to the original data. For illustration purposes, the 97.5$^{\text{th}}$ and the 2.5$^{\text{th}}$ quantiles $Q_{97.5}$, $Q_{2.5}$ respectively corresponding to the 95% confidence region are also presented in Table II, along with the mean quantile-quantile (Q-Q), illustrated in red in Fig. 10. We observed a slight increase in the standard deviation of the forecast errors, but in general the results matched the measured data closely. Similarly, the results in terms of nRMSE were also consistent. Moreover, the CRPS was notably low for all locations, indicating the precision of the probabilistic forecast scenarios. Note that a value of zero corresponds to a perfect deterministic forecast. The quantiles of the simulated forecast errors agreed with the theoretical quantiles of the measured forecast errors during most of the sample domain. However, we observed significant deviations below the 2.5$^{\text{th}}$ quantile in all cases except for West Flanders. In other words, the model tended to amplify those situations in which the the solar generation was under–predicted i.e. the forecast was smaller than the actual realization.

V. Discussion

There are three main reasons behind the limitations described at the end of the previous section. Firstly, the statistical distribution of the forecast errors was not modelled specifically. Normalized Gaussian forecast scenarios are transformed back to the original space via the inverse transformation in eq. (13) assuming that each random variable

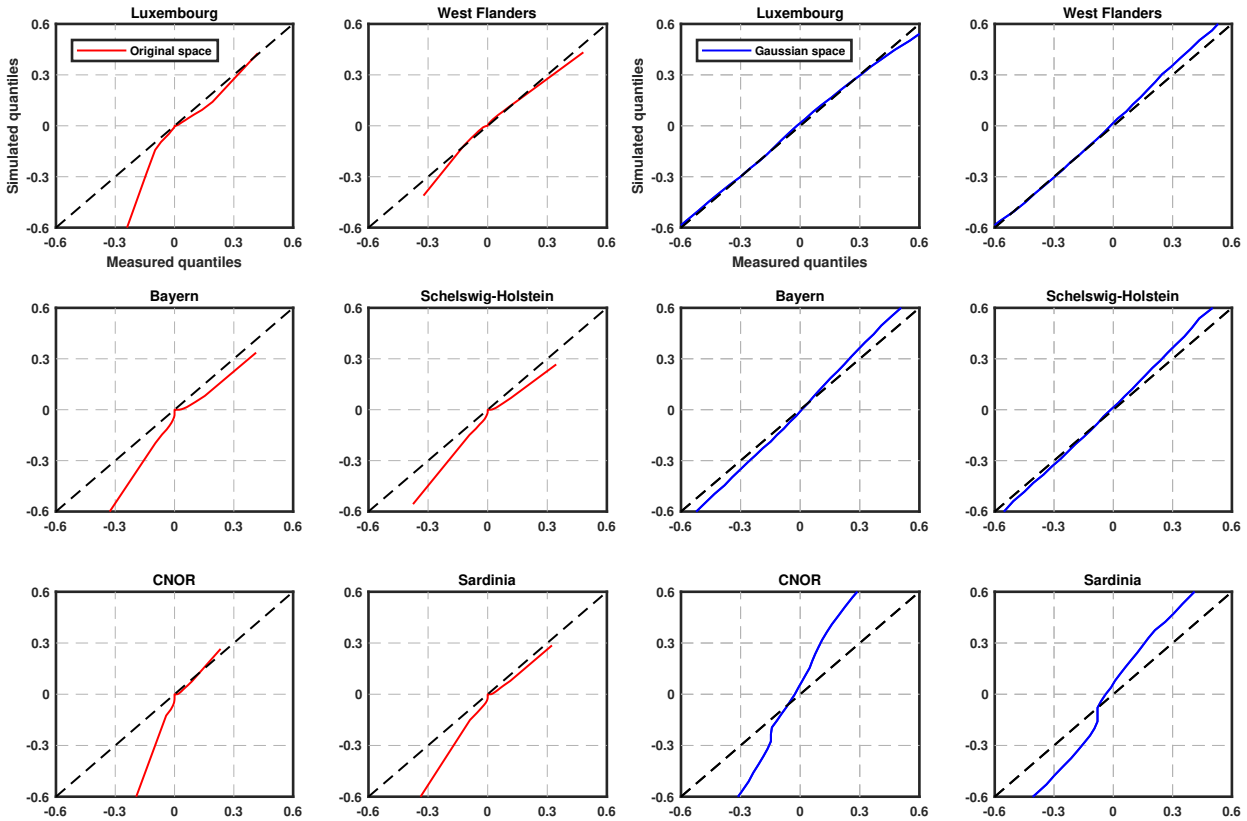

Fig. 10: Mean quantile-quantile plot of actual and simulated forecast scenarios. Red lines correspond to the original domain i.e. $\boldsymbol{\epsilon_t} = \widetilde{\boldsymbol{x}}_t - \boldsymbol{x}_t$ and $\boldsymbol{\epsilon}_t^{sim} = \widetilde{\boldsymbol{x}}_t^{sim} - \boldsymbol{x}_t$, whereas Gaussian vectors are presented in blue i.e. $\boldsymbol{\xi_t}, \boldsymbol{\xi}_t^{sim}$. In both cases, dashed black lines show the theoretical quantiles.

$\widetilde{z}_{i,t}^{sim}$ strictly follows a standard normal distribution $\mathcal{N}(0,1)$. This condition is not fully guaranteed by eq. (12) which may introduce additional errors in the transformed vector. For instance the Gaussianity of the elements of $\boldsymbol{\xi}_{\boldsymbol{t}}^{\boldsymbol{sim}}$ might depend on the considered number of samples. Furthermore, the derivation of the vector $\boldsymbol{z_t}$ requires different steps and it is likely to deviate from a strictly standard multivariate normal distribution. Additional errors are intrinsic to the transformations F_i for $i = 1, \ldots, k$ which are essentially empirical and non–linear; hence limiting the tractability of the problem. The effect of such transformations in the simulated forecast error scenarios can be noticed in Fig. 10, where the blue lines represent average quantiles of the simulated normalized Gaussian forecast errors against the quantiles of the normalized measurements. As expected, the quantiles of the simulated scenarios prior to the transformation are much closer to the quantiles of the measured forecast errors. Last but not least, the clear–sky power time series could be additional sources of inaccuracies, specially due to the estimation of the hourly installed PV capacity in each area, $p_{i,t}$. These issues will be considered in future research in order to improve the accuracy of the model.

An additional limitation of the presented technique is that it only reproduces the long–term (i.e. day–ahead) properties of the forecasts. For example, the model is calibrated for day–ahead simulations, which does not ensure correct intra–day behaviour. However, we believe that the methodology is sound and general enough to be extended to other temporal resolutions e.g. intra-hourly, and/or forecast horizons e.g. intra–day, as long as data are available. The deterministic components of solar radiation become less important as the time resolution increases. Therefore, the data pre–processing step might be avoided, simplifying the procedure. The model can be improved by including off–diagonal terms in the parameter matrices (8). This will require to consider different expressions for each of the cross–correlation lags and can lead our future research efforts. Furthermore, geostatistics offer different alternatives to represent the spatial variation of the forecast cross-correlation coefficients. More precisely, kriging [45] emerges as a sound technique for spatial interpolation. It has already been applied in the wind and solar literature [46, 47] and it is a possibility to be explored in the near future

VI. Conclusion

Solar photovoltaic (PV) power is starting to play a significant role in modern power systems. This fact highlights the importance of improving how solar generation is currently represented in grid planning and operation. Hereof, capturing the uncertainty arising from solar forecast errors represents a big challenge specially when historical forecasts are not available or can potentially become a tedious exercise using NWP–based methods. In this work, we propose a general transformation–based heuristic methodology to generate hourly day–ahead forecast scenarios matching both spatial and temporal characteristics as well as long–term statistical properties of observed data. The temporal properties of the forecast errors, summarized by their autocorrelation function, are related to the geographical spread of the PV panels inside a given area. Besides, cross-correlation properties are parametrized as a function of the distance between areas. Hence, only information regarding the geographical spread of the PV panels and aggregated measurements are required to simulate forecast errors. We observed both strong daily and seasonal patterns in the data at this level of temporal resolution. Therefore, our model relies on a sequence of transformations in order to correct for non–stationarity and a changing error variance. The stochastic modelling takes place in the standard normal space and assumes a VAR(2) process. We presented a case study considering six out–of–sample regions with different sizes and climatic characteristics. Despite the simplicity of the model, the sequence of transformations captured both temporal and spatial properties of area forecasts. Thus, no higher order models or dummy variables were required to retain the daily structure of the forecast errors. In addition, the statistical distributions of the forecast errors were preserved between the 2.5^{th} and 97.5^{th} quantiles. Based on the validation results, we believe that the approach could be applied to areas for which forecast records are not currently available. Furthermore, the parameters of the model can be re-calibrated in order to incorporate advances in solar generation prediction techniques. The proposed methodology could be applied to generate forecast scenarios as an input to different power system studies e.g. focusing on operational reliability and renewable integration. This would allow analysing the impact of PV uncertainty on the power system via stochastic scenarios. Ultimately, the method can be compared against related probabilistic techniques e.g. meteorological ensembles, as well as alternative ways of capturing uncertainty e.g. chance–constrained power flow formulations; paving the way for a more efficient integration of renewable energy sources in the power system.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 608540 GARPUR, NSON-DK (ForskEL) and Flex4RES (Nordic Energy Research) projects.

References

[1] R. H. Inman, H. T. Pedro, and C. F. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in energy and combustion science*, vol. 39, no. 6, pp. 535–576, 2013.

[2] S. Pelland, J. Remund, J. Kleissl, T. Oozeki, and K. De Brabandere, "Photovoltaic and solar forecasting: state of the art," *IEA PVPS, Task*, vol. 14, pp. 1–36, 2013.

[3] E. Lorenz, T. Scheidsteger, J. Hurka, D. Heinemann, and C. Kurz, "Regional pv power prediction for improved grid integration," *Progress in Photovoltaics: Research and Applications*, vol. 19, no. 7, pp. 757–771, 2011.

[4] N. Siebert, *Development of methods for regional wind power forecasting*. PhD thesis, École Nationale Supérieure des Mines de Paris, 2008.

[5] 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, 2015.

[6] P. Pinson, H. Madsen, H. A. Nielsen, G. Papaefthymiou, and B. Klöckl, "From probabilistic forecasts to statistical scenarios of short-term wind power production," *Wind energy*, vol. 12, no. 1, pp. 51–62, 2009.

[7] 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.

[8] J. Jeon and J. W. Taylor, "Using conditional kernel density estimation for wind power density forecasting," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 66–79, 2012.

[9] N. Zhang, C. Kang, Q. Xia, and J. Liang, "Modeling conditional forecast error for wind power in generation scheduling," *IEEE Transactions on Power Systems*, vol. 29, no. 3, pp. 1316–1324, 2014.

[10] A. Staid, J.-P. Watson, R. J.-B. Wets, and D. L. Woodruff, "Generating short-term probabilistic wind power scenarios via nonparametric forecast error density estimators," *Wind Energy*, 2017.

[11] 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, 2016.

[12] H. Holttinen, L. Söder, E. Ela, *et al.*, *Design and operation of power systems with large amounts of wind power: Final report, IEA WIND Task 25, Phase one 2006-2008*. VTT Technical Research Centre of Finland, 2009.

[13] L. Soder, "Simulation of wind speed forecast errors for operation planning of multiarea power systems," in *Probabilistic Methods Applied to Power Systems, 2004 International Conference on*, pp. 723–728, IEEE, 2004.

[14] P. Meibom, R. Barth, B. Hasche, H. Brand, C. Weber, and M. O'Malley, "Stochastic optimization model to study the operational impacts of high wind penetrations in ireland," *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 1367–1379, 2011.

[15] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Solar Energy*, vol. 83, no. 10, pp. 1772–1783, 2009.

[16] R. Bessa, A. Trindade, C. S. 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, 2015.

[17] 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, 2015.

[18] E. Nuño, M. Koivisto, N. Cutululis, and P. Sørensen, "Simulation of regional day-ahead PV power forecast scenarios," in *PowerTech, 2017 IEEE Manchester*, IEEE, In Press, 2017.

[19] T. Huld and A. M. G. Amillo, "Estimating pv module performance over large geographical regions: The role of irradiance, air temperature, wind speed and solar spectrum," *Energies*, vol. 8, no. 6, pp. 5159–5181, 2015.

[20] J. Ekstrom, M. Koivisto, I. Mellin, J. Millar, and M. Lehtonen, "A statistical model for hourly large-scale wind and photovoltaic generation in new locations," *IEEE Transactions on Sustainable Energy*, 2017.

[21] B. Klöckl and G. Papaefthymiou, "Multivariate time series models for studies on stochastic generators in power systems," *Electric Power Systems Research*, vol. 80, no. 3, pp. 265–276, 2010.

[22] U. Focken, M. Lange, K. Mönnich, H.-P. Waldl, H. G. Beyer, and A. Luig, "Short-term prediction of the aggregated power output of wind farmsa statistical analysis of the reduction of the prediction error by spatial smoothing effects," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 90, no. 3, pp. 231–246, 2002.

[23] P. Norgaard and H. Holttinen, "A multi-turbine power curve approach," in *Nordic wind power conference*, vol. 1, pp. 1–2, 2004.

[24] E. Nuño and N. Cutululis, "Generation of large-scale PV scenarios using aggregated power curves," in *Power and Energy Society General Meeting, 2017 IEEE*, IEEE, In Press, 2017.

[25] A. N. Hahmann, D. Rostkier-Edelstein, T. T. Warner, F. Vandenberghe, Y. Liu, R. Babarsky, and S. P. Swerdlin, "A reanalysis system for the generation of mesoscale climatographies," *Journal of Applied Meteorology and Climatology*, vol. 49, no. 5, pp. 954–972, 2010.

[26] E. Nuño, P. Maule, A. Hahmann, N. Cutululis, P. Sørensen, and I. Karagali, "Simulation of transcontinental wind and solar PV generation time series," *Renewable Energy (In Press)*.

[27] J. Ekström, M. Koivisto, J. Millar, I. Mellin, and M. Lehtonen, "A statistical approach for hourly photovoltaic power generation modeling with generation locations without measured data," *Solar Energy*, vol. 132, pp. 173–187, 2016.

[28] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*. Springer Science & Business Media, 2013.

[29] E. Lorenz, J. Hurka, D. Heinemann, and H. G. Beyer, "Irradiance forecasting for the power prediction of grid-connected photovoltaic systems," *IEEE Journal of selected topics in applied earth observations and remote sensing*, vol. 2, no. 1, pp. 2–10, 2009.

[30] A. Boone, *Simulation of short-term wind speed forecast errors using a multi-variate ARMA (1, 1) time-series model*. Master Thesis, Royal Institute of Technology, Sweden, 2005.

[31] N. J. Higham, "Computing a nearest symmetric positive semidefinite matrix," *Linear algebra and its applications*, vol. 103, pp. 103–118, 1988.

[32] Elia, "Solar-PV power generation data." http://www.elia.be/en/grid-data/power-generation/Solar-power-generation-data/Graph. [Online; accessed 15-November-2017].

[33] Amprion, "Photovoltaic-Infeed." https://www.amprion.net/Grid-Data/Photovoltaic-Infeed/. [Online; accessed 15-November-2017].

[34] TenneT, "Actual and forecast photovoltaic energy feed-in." https://www.tennettso.de/site/en/Transparency/publications/network-figures/actual-and-forecast-photovoltaic-energy-feed-in. [Online; accessed 15-November-2017].

[35] TransnetBW, "Grid Data." https://www.transnetbw.com/en/transparency/market-data/key-figures. [Online; accessed 19-July-2008].

[36] ENTSO-E, "Transparency Platform." https://transparency.entsoe.eu/. [Online; accessed 15-November-2017].

[37] Bundesnetzagentur. https://www.bundesnetzagentur.de. [Online; accessed 15-November-2017].

[38] GSE, "Atlasole." http://atlasole.gse.it/atlasole/. [Online; accessed 15-November-2017].

[39] Commission wallonne pour l'Energie. http://www.cwape.be/. [Online; accessed 15-November-2017].

[40] VREG. http://www.vreg.be/nl/groene-stroom. [Online; accessed 15-November-2017].

[41] W. Skamarock, "Coauthors, 2008: A description of the advanced research WRF version 3. NCAR Tech. Note," tech. rep., NCAR/TN-475+STR.

[42] B. Haurwitz, "Insolation in relation to cloudiness and cloud density," *Journal of meteorology*, vol. 2, no. 3, pp. 154–166, 1945.

[43] R. Perez, P. Ineichen, R. Seals, J. Michalsky, and R. Stewart, "Modeling daylight availability and irradiance components from direct and global irradiance," *Solar energy*, vol. 44, no. 5, pp. 271–289, 1990.

[44] J. E. Matheson and R. L. Winkler, "Scoring rules for continuous probability distributions," *Management science*, vol. 22, no. 10, pp. 1087–1096, 1976.

[45] N. Cressie, "Spatial prediction and ordinary kriging," *Mathematical geology*, vol. 20, no. 4, pp. 405–421, 1988.

[46] M. Cellura, G. Cirrincione, A. Marvuglia, and A. Miraoui, "Wind speed spatial estimation for energy planning in sicily: A neural kriging application," *Renewable energy*, vol. 33, no. 6, pp. 1251–1266, 2008.

[47] 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, 2013.