Citation	Mahtab Kaffash, Kenneth Bruninx, Geert Deconinck, (2021), Data-driven forecasting of local PV generation for stochastic PV-battery system management International Journal of Energy Research, 9 May 2021
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	https://onlinelibrary.wiley.com/doi/full/10.1002/er.6826
Journal homepage	https://onlinelibrary.wiley.com/journal/1099114x
Author contact	Mahtab.kaffash@kuleuven.be phone number + 32 (0)16 377888

(article begins on next page)

ARTICLE TYPE

Data-driven forecasting of local PV generation for stochastic PV-battery system management

Mahtab Kaffash^{*1,2} | Kenneth Bruninx^{2,3} | Geert Deconinck^{1,2}

¹ESAT-Electa, KU Leuven, Kasteelpark
Arenberg 10 bus 2445, 3001 Leuven,
Belgium

²EnergyVille, Thor Park, 8310, 3600 Genk,
Belgium

³WTK-TME, KU Leuven, Celestijnenlaan
300, 3001 Leuven, Belgium

Correspondence

*Mahtab Kaffash. Email:
mahtab.kaffash@kuleuven.be

Present Address

ESAT-Electa, KU Leuven, Kasteelpark
Arenberg 10 bus 2445, 3001 Leuven,
Belgium

Summary

Power systems face more uncertainty by increasing photovoltaic system installations on the roof of buildings. To optimally manage energy and available flexibility in a building, stochastic optimization is used to take an optimal decision under uncertainty and minimize the operational cost. In stochastic optimization, a scenario set is used as an input to represent the uncertainty in a random variable, PV energy generation in this case. In this paper, a data-driven method is proposed to obtain the distribution of the random variable and later generate scenario sets representing the uncertainty on day-ahead PV energy generation installed on the roof of a building. This method is only based on historical PV generation and it does not require any other external data such as weather forecasts. A machine learning-based technique is applied to forecast the PV energy production following by generating a scenario set for day-ahead decision making. Later, the day-ahead PV-battery system management problem is formulated as a two-stage stochastic optimization while the generated scenario set is the input of this optimization. The proposed algorithm is tested in the day-ahead scheduling of a PV-battery system for a commercial building, informed by real-life measurement data. The results show that the proposed algorithm is able to capture the uncertainty in PV system while providing a cost-optimal and reliable solution to the application problem. Without using any weather data, the error of the proposed PV energy forecast method reaches to NRMSE=11.89, while there is 5% reduction in the operational cost of the proposed two-stage stochastic optimization. Moreover, the proposed algorithm is easy to implement in the energy management system of a building to manage PV-battery system.

KEYWORDS:

Data-driven forecasting, Non-parametric distribution, PV-battery system, Stochastic optimization, Uncertainty of PV system

1 | INTRODUCTION

The past decade has witnessed a radical increase in the capacity of photovoltaic (PV) system, installed on the roof of buildings. It is also predicted that by 2050 PV systems will see the largest capacity increase of all electricity generation technologies¹.

Nomenclature

Indices

k	breakpoints of piecewise function
s	scenario of PV energy generation
t	time step

Parameters

Δt	quarter of hour
η_b	efficiency of the battery (%)
C^b	capacity of the battery (kWh)
C^{da}	day-ahead electricity price (€/kWh)
$C^{pur,rt}$	real-time purchased electricity price (€/kWh)
$C^{sold,rt}$	real-time sold electricity price (€/kWh)
E^l	demand of the building (kWh)
E^{PV}	PV energy generation (kWh)
K	number of breakpoints of piecewise function
L	breakpoints of ΔSoE
N	number of scenarios in out-of-sample analysis
$P^{ch,max}$	maximum charging power (kW)
$P^{dis,max}$	maximum discharging power (kW)

$Prob$	probability of scenario
R	breakpoints of SoE
S	total number of scenarios
SoE^{min}	minimum state-of-energy (kWh)
SoE_0	initial state-of-energy (kWh)
T	forecast horizon
T'	optimization horizon

Variables

γ	binary variable of the battery
$e^{ch,b}$	charging energy of the battery (kWh)
$e^{dis,b}$	discharging energy of the battery (kWh)
$e^{pur,da}$	day-ahead purchased energy (kWh)
$e^{pur,rt}$	real-time purchased energy (kWh)
$e^{sold,da}$	day-ahead sold energy (kWh)
$e^{sold,rt}$	real-time sold energy (kWh)
obj	value of objective function (€)
soe	battery state-of-energy (kWh)

This significant growth of PV capacity reduces the dependency on fossil fuels to generate electricity and avoids greenhouse gas emission as well.

On the other hand, since the output of PV system is not controllable, there is always a mismatch between local PV electricity generation and the energy demand of the building. A battery can be used together with a PV system in the building to not only compensate the mismatch between demand and supply, but also to reduce the operational cost of the building by providing flexibility. The battery can store the surplus PV generation and supply the demand when there is less (or no) generation². The battery can also charge during periods of low electricity price and then discharge to supply the demand when the price of electricity is high to reduce the operational cost³. In fact, the optimal schedule of the battery depends on the future local PV generation. However, the power produced by PV is influenced by solar radiation and weather conditions which make the PV output highly variable and limitedly predictable. The limited predictability of PV systems mounted on the roofs of buildings bring a lot of challenges in the scheduling and operation of PV-battery system⁴. Especially, in the future these systems are expected to be more exposed to the market⁵, as this will trigger the need for day-ahead decision making.

As discussed above, the optimal day-ahead energy management of the PV-battery system depends on the actual PV energy generation. Since this PV generation is uncertain during day-ahead decision making, a scenario set can be generated to represent possible PV generation in the future. In the literature, a large number of studies have been conducted on the optimal energy management in a PV-battery system, from deterministic approach to stochastic approach. All of these approaches rely on the estimation of the PV generation in the future. Therefore, depending on the way that PV energy generation is represented, the optimal decision making approach will be selected.

In deterministic approach, the generation of PV system in the future is shown by a single signal. The very basic method to represent PV energy generation in the future is the persistence technique, which means that the output of PV system at each time step in next day is the same as today⁶. However, the results of this technique can be accurate only for lead-time 4-6 hours⁶. Persistence technique was used in⁷ to manage PV-battery system. In time series forecasting, statistical methods, such

as ARIMA, were proposed, which are based on historical observations. For instance, ARIMA was selected in⁸ to forecast PV energy generation. Then, the forecast time series was used to manage flexibility provided by the battery of electric vehicles. Although, statistical methods are simple and fast, they are not usually able to capture the dynamic of the complex systems, such as non-linear behavior and rapid variation⁹, which are both presented in PV generation. In the last decade, the recent developments in machine learning on the one hand and developments in smart meters to provide real data on the other hand make machine learning techniques quite popular for time series point forecast. Therefore, many researchers have focused on point forecast of PV generation^{10,11,4, 12}. The point forecast provides the expected PV generation for each time step over the time horizon. Azuatalam *et al.* proposed a PV energy forecaster based on artificial neural network and used the forecast values to optimally schedule the battery¹³. Although this approach is easy to implement, it leads to unrealistic and inaccurate results because there is always an error in the PV forecast. Therefore, only using the point forecast values to schedule and manage the flexibility ignores the uncertainty in the forecaster. In other words, there is a bias in the decisions towards the point forecast, which may lead to a low real-time performance. As a result, the main disadvantage of deterministic approach is that the uncertainty of PV system is ignored in decision making process.

In recent years, more researchers have been interested in probabilistic forecast. The main goal of this approach is to give information to the decision maker about the uncertainty of the point forecast and the stochastic behavior of random variable. In general, there are two options to obtain the cumulative distribution function (CDF) of a random variable: (i) parametric approach, in which an assumption on the type of the distribution of a random variable is required; (ii) non-parametric approach in which the quantile function is required to estimate the quantiles of the random variable. In the parametric approach, it is assumed that the error of the point forecast is following a pre-defined full parametric distribution, such as Normal distribution^{14,15}, or Beta distribution¹⁶, and then the parameters of that particular distribution are calculated from the data set (check^{17,18} for more details). The second option to obtain the distribution of the random variable is called non-parametric approach, in which quantile function is calculated. In existing literature, several methods have been proposed to deal with probabilistic PV energy generation forecast^{19,20}. Moreover, the application of probabilistic forecast of PV generation in scheduling the battery and managing flexibility of the system was studied in the literature²¹. Although probabilistic forecast provides some information about the stochastic behavior of random variable, it considers the output of the PV system on each time step as a uncorrelated random variable, hence, ignoring the auto-correlation in PV output²². For instance, to manage the day-ahead flexibility, we are facing with a multi-stage decision making problem in which the interdependence structure of forecast error matters.

As a solution, scenario generation approach was proposed to deal with multi-stage decision making in the presence of random variable²³. Therefore, stochastic optimization is used to make optimal decision when PV energy generation is shown by scenarios. A set of scenarios is the discrete representation of random variable, which includes the temporal dependency of the random variable over the time horizon of the problem, which is ignored in probabilistic forecast. In order to generate scenarios, the full probability distribution of random variable is required. Later on, the scenario set is generated by randomly sampling from the distribution. Conventionally, in the existing literature, to generate scenario set for wind power and PV power generation, a parametric distribution was used to model the distribution of forecast error and the parameters of that particular distribution are calculated from historical data set. In¹⁷, a t location-scale distribution was assumed for the distribution of the wind power to generate scenarios. A normal distribution was assumed in^{24, 25} to generate scenarios for wind and PV systems to stochastically schedule a microgrid under uncertainty of wind and PV.

In all above literature, an assumption on the distribution model of the forecast error is required to generate scenarios to deal with uncertainty of PV power generation. In fact, it is not always easy to find the best distribution which can explain the features of the available data set.

In contrast to parametric approach, In 2009, Pinson *et al.* proposed a method to generate scenarios for wind farm from a non-parametric distribution, obtained from the error of point forecast for wind power generation²³. Thanks to this proposed methodology, there has been no need for assumptions on the distribution of the forecast error and scenarios can be generated from non-parametric distribution. Later, the same method to generate scenarios was applied to PV systems in^{26,27}.

Since the generated scenario set affects the performance of the stochastic optimization, it is important to evaluate the quality of generated scenario set by the provided optimal decision to our problem²⁸.

Unlike large PV installation, for the scale of one building, there is more variation in the output of PV system, plus the weather forecast and PV forecast data are not freely available to the energy management system of the building. Unfortunately, there has been relatively little analysis in the literature about modeling the uncertainty in a single PV system installed on a building. In most of the aforementioned literature, the focus was on either design a machine learning based point forecast by using both historical generation data and weather forecast, or describing the uncertainty of the large-scale PV installation while the PV

FIGURE 1 Overview of the proposed method

point forecast data was provided by another party. Therefore, in this paper we propose a machine learning based point forecast followed by a scenario generation method, which is completely independent of the weather forecast data, in order to capture the rapid variation in the output of the PV system. The proposed algorithm can be easily implemented in the building energy management system.

In this paper, we address the challenge of generating a scenario set capturing the uncertainty and auto-correlation in the output of a single PV system installed on the roof of a building. We leverage a data-driven technique to generate these scenarios based on historical measurements alone. Figure 1 shows the overview of the proposed method in this paper. First, a machine learning based method is introduced in Section 2 to provide point forecast. This step is followed by a non-parametric method to obtain the predictive distribution (probabilistic forecast) of the PV generation and to translate this to scenarios. Second, a two-stage stochastic optimization is proposed in Section 3 to obtain the optimal schedule of the PV-battery system. The main contribution of this paper is to propose a data-driven PV scenario generation tool that allows capturing diurnal and seasonal variations of the system while only relies on the local historical PV measurements. Moreover, the quality of the generated scenario sets is demonstrated in the stochastic PV-battery schedule on the real-life data set. Therefore, the main objectives of this work can be listed as below:

- Propose a point forecast method for PV energy generation which only depends on historical PV generation data. The accuracy of this method should be compared with other popular PV forecast methods that require weather data to predict PV generation.
- Build a reliable non-parametric PV forecaster in which no assumption is needed on the distribution of PV generation.
- Generate scenario set independent of the weather data, which is representing the uncertainty of PV energy production.
- Linearly formulate system, including PV and battery as a two-stage stochastic optimization
- Evaluate the quality of generated scenario set not only by some multivariate evaluation metrics, but also by analysing the optimal solution of stochastic optimization.

2 | DATA-DRIVEN PV SCENARIO GENERATION

As mentioned before, scenarios are discrete realizations of the random variable. In order to generate scenarios, one needs the information about the distribution of the random variable. Algorithm 1 shows an overview of the proposed algorithm to forecast the distribution and generate scenario set for PV system. According to Algorithm 1, assume that the historical observed PV energy generation data is given. The first step of the proposed algorithm is to develop day-ahead point forecast. However, as a pre-processing step, we propose to apply feature selection to detect important lags and improve the performance of point forecast. Later on, the predictive marginal distribution function is calculated for each time step in a day-ahead forecast horizon. Then, the multivariate distribution function, describing the error on this forecast, is obtained as a Gaussian Copula based on the predictive marginal distribution functions. Finally, by randomly sampling from obtained multivariate distribution function, the scenarios are generated by superimposing them on the forecast.

2.1 | Feature selection

As mentioned before, the focus of this paper is to model and predict the uncertainty of PV energy generation installed on a single building, where the weather forecast information and PV forecast values are not freely available to the building energy management system. The only available information is historical PV energy generation data measured by a local measurement system. As the result, in order to build the algorithm to obtain the point forecast, selecting the important lags for each time step is necessary to improve the efficiency of the forecast method. Especially for PV system, due to diurnal and seasonal behavior (higher generation in the middle of the day and during summer), feature selection and removing non-relevant lags can significantly improve the performance of a data-driven forecaster. As a result, including the measured data for different hours of the day and different seasons to build forecaster lead to decrease the performance of the forecast. For instance, if the target time is in summer, including the data of the winter will increase the error of the forecast. In this paper, least absolute shrinkage and selection operator (LASSO), a data-driven method, is used to select the relevant features.

In general, LASSO is able to deal with spatio-temporal dependency (in this paper, only temporal dependency matters since the focus is on a single location)²⁹. LASSO can be used to decrease the number of features and obtain relevant lags. Let $p_d(t)$ be the feature vector (historical PV generation) for day d at time t . The PV generation forecasting problem for the next day ($d + 1$) at time step t can be modelled as a linear regression problem:

$$\hat{p}_{d+1}(t) = \hat{\beta}_0 + \hat{\beta}_1 p_d(t) + \dots + \hat{\beta}_l p_{d-l+1}(t) \quad (1)$$

where l is the maximum lag and β s are the coefficients. There are different ways to estimate β values and LASSO is one of them. LASSO does not only avoid overfitting, but is also able to obtain relevant features by assigning zero to the coefficient(s) of irrelevant inputs³⁰. Let $P(t) = [P_1(t), P_2(t), \dots, P_D(t)]^{D \times l}$ and $p(t) = [p_1(t), p_2(t), \dots, p_D(t)]^D$, whereas D is the total number of days, p_i is the real measurement and P_i includes l features for each observation. In LASSO, one can estimate the optimal values for coefficients by solving the following problem:

$$\hat{\beta} = \arg \min_{\beta} \|p - P^T \beta\|^2 \text{ s.t. } \|\beta\|_1 \leq \eta_{LASSO} \quad (2)$$

Algorithm 1 : PV scenario generation algorithm

Input: real PV generation data (P), number of scenarios (S)

- 1: Apply LASSO on data set to obtain optimal lags, l , β s in (1)
 - 2: Obtain day-ahead point forecast ($\hat{P}$) by FFNN while its input dimension is based on l , β s
 - 3: Calculate the quantiles of the error based on selected features
 - 4: Obtain the CDF of the random variable, $\hat{F}(P)$
 - 5: Transfer data to multivariate Gaussian variable while $Z \sim \mathcal{MVN}(0, \Sigma)$, using PIT
 - 6: Calculate empirical covariance matrix based on l
 - 7: Generate S random values respecting $\mathcal{MVN}(0, \Sigma)$
 - 8: Transfer generated scenarios from multivariate Gaussian space to original random space
 - 9: Calculate S scenarios for day-ahead PV generation according to the generated scenarios for the forecast error
-

FIGURE 2 A feed-forward neural network diagram to forecast day-ahead PV energy generation

where η_{LASSO} is LASSO positive tuning parameter. Finally, by solving (2), one can obtain the optimal value for l , as well as β values (only non-zero values matter) for each time step of the next day. In order to evaluate the performance of LASSO, the normalized root-mean square error (NRMSE) is used³¹, which is defined by:

$$NRMSE = \frac{\sqrt{\frac{1}{D \times T} \sum_{d=1}^D \sum_{t=1}^T (\hat{p}_d(t) - p_d(t))^2}}{P_{max} - P_{min}} \quad (3)$$

where P_{max} and P_{min} are the maximum and minimum values of observation, which are here set to the total capacity of inverters and zero, respectively. In other words, NRMSE is calculated by dividing RMSE, which is a popular metric to evaluate the performance of forecast method, by the installed capacity of PV system. Since NRMSE is reported in %, it makes possible to compare the results of proposed PV forecast with other papers.

2.2 | Obtain point forecast

A feed-forward neural network (FFNN) is applied on data set to forecast the day-ahead PV energy generation at each time of the day. Figure 2 illustrates the diagram of proposed method, where the inputs are the relevant lags obtained from (2) and the outputs are the day-ahead prediction values (each neuron corresponds to one time step). This means that the number of neurons at input layer and output layer are fixed. However, in order to have the best architecture for the proposed FFNN, the hyper-parameters, including the number of hidden layers, the number of neurons in each hidden layer, and the activation function in each layer should be tuned during the learning process to reach to the lowest error. Grid search is used to find the optimal hyper-parameters for the FFNN. NRMSE is selected as the performance metric.

The available data set is divided into three categories: training, validation, and test set. The hyper-parameters of the proposed method are defined on the training set. The validation set allows avoiding overfitting. Finally, the optimal architecture of FFNN is applied on the test set to evaluate the accuracy of the selected model.

2.3 | Obtain the predictive distribution function

Let random variable $\mathcal{P}$ and its CDF ($F(\mathcal{P})$), quantile q^α of $\mathcal{P}$ is the value at which the probability of $\mathcal{P}$ is less than or equal to α , i.e.:

$$\alpha = F(q^\alpha) \quad (4)$$

where α is the probability value between $[0, 1]$. According to (4), the CDF of the random variable $\mathcal{P}$ can be determined via quantiles.

In this paper, empirical quantile method is used to obtain the quantiles of predictive CDF. This method is based on the quantiles of the error of historical point forecast. Let assume that the historical PV production, as well as corresponding point forecasts

for D days for all time steps are available, $([p_1(t), \dots, p_D(t)])$ and $([\hat{p}_1(t), \dots, \hat{p}_D(t)])$, respectively. Then, the historical error can be computed via:

$$e_d(t) = p_d(t) - \hat{p}_d(t) \quad \forall t, \forall d \quad (5)$$

For each time step, the historical error can be calculated via (5) and then the empirical quantile distribution of the errors ($q^\alpha(e)$) can be calculated as a proxy for the CDF of the forecast error. Moreover, it is shown in³² that LASSO is able to improve the performance of probabilistic forecast if it used as feature selection. In this paper, in order to increase the accuracy of our estimate of the CDF of the forecast error, for each day and for each time step, only the errors of same features (explained in Section 2.1) are included to build the empirical quantile function[†]. This means the empirical quantile function for $d + 1$ at time step t ($q^\alpha(e_{d+1}(t))$) is obtained based on $e_d(t), e_{d-1}(t), \dots, e_{d-l+1}(t)$. Afterwards, one can obtain the quantile function of PV generation for day-ahead ($d + 1$) at time step t as:

$$q^\alpha(p_{d+1}(t)) = \hat{p}_{d+1}(t) + q^\alpha(e_{d+1}(t)) \quad \forall t \quad (6)$$

Finally, the empirical marginal CDF of PV generation ($\hat{F}_t$) for each time step t can be obtained via:

$$\alpha = \hat{F}_t(q^\alpha(p_{d+1}(t))) \quad (7)$$

To evaluate the performance of the probabilistic forecast, reliability (calibration) is analyzed in this paper. According to²⁰, reliability is an important metric for probabilistic forecast. More evaluation metrics can be found in³³, however this discussion is out of the scope of this paper. It should be clear out that the focus of this paper is on forecasting the distribution of random variable, not building the distribution around the point forecast as it was done in³⁴.

2.4 | Multivariate distribution

The individual PV predictive marginal distribution for day-ahead for each time step can be obtained as explained above. In order to obtain the joint distribution over the forecast horizon T , a copula function C can be defined to transform individual CDF of each random variable to a multivariate distribution function³⁵:

$$H(p_1, p_2, \dots, p_T) = C(\hat{F}_1(p_1), \hat{F}_2(p_2), \dots, \hat{F}_T(p_T)) \quad (8)$$

where H is the predictive multivariate distribution function. Based on the definition of copula, H is a multivariate cumulative distribution function for which the marginal cumulative probability distribution of the individual variable is uniform on the interval $[0, 1]$. Following²⁶, Gaussian copula is used to represent the multivariate distribution of random variable $\mathcal{P}$. As a result, in order to find the multivariate distribution function, the random variable $\mathcal{P}_t(t = 1, 2, \dots, T)$ should be transformed to a standard uniform distribution. Let's define a new variable Y_t , where:

$$Y_t = \hat{F}_t(p_t) \quad \forall t \quad (9)$$

based on Probability Integral Transformation (PIT), Y is following a standard uniform distribution ($Y \sim U[0, 1]$)³⁶.

A new random variable X is introduced, such that X is following a multivariate Gaussian distribution, $X \sim \mathcal{MVN}(\mu, \Sigma)$. Then we can write the following equation:

$$X = \sqrt{2} \operatorname{erf}^{-1}(2Y - 1) \quad (10)$$

where erf^{-1} is the inverse of the error function. It should be pointed out that PIT also implies that the CDF of X is uniform in $[0, 1]$ as well, so that it is possible to use the transformation process described in²³.

The covariance matrix (Σ), with the dimension $T \times T$, represents the interdependence of the multivariate Gaussian distribution variable over the forecast horizon T . Since the real measurements are available, the covariance matrix is calculated empirically in this paper, as suggested in²⁷. However, due to diurnal and seasonal behavior of the PV generation, in order to calculate the empirical covariance matrix only the historical data until lag l , obtained from LASSO is used. It should be mentioned that it is needed to obtain the covariance matrix on transformed data, rather than the raw data (which doesn't follow the multivariate Gaussian distribution).

[†]Due to the diurnal and seasonal behavior of the PV system, there is more production during the day and during summer. Consequently, the absolute value of the forecast error is also higher in the middle of the day and during the summer time (comparing to the winter time).

2.5 | Scenario generation

In order to generate scenario set, including S realizations of the possible day-ahead PV generation, the steps below should be followed:

- Transform from random variable domain to multivariate Gaussian domain according to²³, as well as calculate Σ based on historical data until lag l .
- Generate S samples from $\mathcal{MVN}(\mu, \Sigma)$. It should be noted that each sample is a vector of length T and the samples are equiprobable.
- Transform the generated samples from multivariate Gaussian to uniform distribution domain by:

$$y_{s,t} = \Phi(x_{s,t}) \quad \forall s, \forall t \quad (11)$$

where Φ is the inverse probit function.

- Transform uniform variable $y_{s,t}$ from uniform domain to original random variable domain and generate the possible scenarios for day-ahead PV generation by:

$$\hat{p}_{s,t} = \hat{F}_t^{-1}(y_{s,t}) \quad \forall s, \forall t \quad (12)$$

2.6 | Evaluation of the generated scenario set

According to the above steps, S scenarios can be generated for the day-ahead PV generation. In fact, the main goal of any scenario generation method is to generate scenarios which are an adequate representation of the random variable¹⁸. Different metrics to evaluate the quality of generated scenarios for the application of wind and PV system were introduced in^{37,26}. In this paper only two popular metrics (according to³⁸) have been used: Energy score (ES) and Variogram-based score (VS) (see appendix A for formulation and check²⁶ for more explanation). Both metrics are proper and negatively-oriented, meaning that a lower score indicates a better performance.

However, since these scenarios are the inputs of the stochastic programming, the quality of the generated scenarios should be evaluated based on the optimal solution of stochastic optimization²⁸. Following^{18,39}, we will test stability and bias of the solution. The stability is evaluated by two analyses: in-sample and out-of-sample analyses. In-sample analysis refers to generating limited number of scenarios, while out-of-sample analysis refers to generating a large number of scenarios and use the optimal values of first-stage variables. If in-sample and out-of-sample stability are reached while also provide an unbiased solution, we can say that the scenario generation method is able to generate enough scenarios to represent the uncertainty of random variable. In addition, in order to ensure comparisons between different model runs are meaningful, we will compute confidence intervals on the out-of-sample objective value as follows¹⁸:

$$[obj - \frac{z_{\alpha/2}\sigma}{\sqrt{N}}, obj + \frac{z_{\alpha/2}\sigma}{\sqrt{N}}] \quad (13)$$

where $z_{\alpha/2}$ is the corresponding score so that the cumulative standard normal probability distribution is $1 - \alpha/2$, σ is the sample standard deviation estimator, and N is the number of scenarios in out-of-sample analysis.

3 | FORMULATION OF THE OPTIMAL PV-BATTERY SYSTEM PROBLEM

In this section, a two-stage stochastic scenario-based optimization is proposed to obtain the optimal schedule of the battery as well as optimal interaction with main grid. The first-stage of the optimization problem deals with day-ahead horizon in which the decision variables are day-ahead purchased and sold energy from and to main grid. These two variables are called here-and-now in stochastic optimization, which are fixed before the realization of random variable. The second-stage of proposed optimization problem deals with real-time scheduling of the battery and real-time interaction with the grid. The variables of the second-stage are called wait-and-see variables, which are scenario-specific variables.

It should be noted that the purpose of this paper is to address the challenge of uncertainty in the PV output forecast installed in a smart (office and industrial) building. A battery is considered as the available flexible asset in this study. However, this proposed method can be adopted to other flexible loads, such as electrical vehicles or heat pumps. Moreover, in order to simplify the problem, the uncertainty of day-ahead electricity price and load are ignored.

3.1 | Objective function

The defined objective function minimizes the expected operational cost of the building over the next day with the resolution of 15 minutes, which ends at $T' = 96$ time steps for each day (the production of PV is assumed zero during night-time). In this paper, the decision variables are the optimal operation of the battery, as well as optimal interaction of the building with the grid, while different scenarios are used to model the uncertainty of the PV generation in the building. The operational cost of the building is formulated in (14) and as it can be seen there are two parts in the defined objective function: the first part represents the operational cost for the first-stage of the optimization problem which is day-ahead operational cost, and the second part corresponds to the real-time operational cost which is modelled in second-stage of optimization problem. In both day-ahead and real-time operation, the energy can be purchased from main grid to supply the demand or charge the battery, or the energy can be sold to the grid. Moreover, a markup is assumed for the real-time interaction to make a difference between day-ahead and real-time electricity price.

$$obj = \min \sum_{t=1}^{T'} C^{da}(t)(e^{pur,da}(t) - e^{sold,da}(t)) + \sum_{s=1}^S Prob(s) \left[\sum_{t=1}^{T'} (e^{pur,rt}(s,t)C^{pur,rt}(t) - e^{sold,rt}(s,t)C^{sold,rt}(t)) \right] \quad (14)$$

It should be mentioned that the production cost of PV is assumed zero and possible revenue via green certificate is not taken into account. It is also assumed the consumer is only a price-taker and does not participate in day-ahead market. It is noted that the optimal sizing of the components is out of the scope of this paper.

3.2 | Constraints

In order to keep the energy balance of the system, the following constraint is defined:

$$e^{sold,rt}(s,t) + e^{sold,da}(t) + E^l(t) + e^{ch,b}(s,t) = e^{pur,rt}(s,t) + e^{pur,da}(t) + E^{PV}(s,t) + e^{dis,b}(s,t) \quad \forall s, \forall t \quad (15)$$

For any PV scenario at each time step, the total energy generated by PV system can be consumed locally, or can charge the battery and the surplus energy should be sold to main grid (in day-ahead and real-time). Moreover, the demand of the building can be supplied by different sources: the main grid, PV system and the battery.

In order to model the battery, the model proposed in⁴⁰ is used. As a result, the battery is model by its state-of-energy (SoE), which is constrained by:

$$soe(s,t) = soe(s,t-1) + \eta_b e^{ch,b}(s,t) - e^{dis,b}(s,t) \quad \forall s, \forall t \quad (16)$$

For any scenario at any time step, the SoE depends on the SoE at previous time step, the amount of charging or discharging at that time step Eq.(16). For the first time step, $t = 1$, the $SoE(s,t-1)$ is equal to the initial SoE (SoE_0). Moreover, the self-discharge of the battery is ignored.

The SoE of the battery should stay in a certain range (Eq.17). This constraint prevents from over-charging/over-discharging and the battery stops interaction when the SoE reaches to the upper-bound/lower-band value.

$$SoE^{min} \leq soe(s,t) \leq C^b \quad \forall s, \forall t \quad (17)$$

To avoid simultaneously charging and discharging of the battery, a binary variable γ is defined. Moreover, there is a limit on the transferred power to charge and discharge the battery. It means that even if the battery is empty/full, there is always a limit to transfer power to charge/discharge it. Defining constraints (18)–(19) deal with this issue.

$$0 \leq e^{ch,b}(s,t) \leq P^{ch,max} \Delta t \gamma(s,t) \quad \forall s, \forall t \quad (18)$$

$$0 \leq e^{dis,b}(s,t) \leq P^{dis,max} \Delta t (1 - \gamma(s,t)) \quad \forall s, \forall t \quad (19)$$

In the above constraints, the battery model is formulated with a constant charging energy limit. However, according to battery charging characteristic, this charging energy limit is not constant in reality. As a result, battery charging capacity ($\Delta SoE(s,t)$) is used to make the model more accurate. ΔSoE means how much energy can be absorbed by the battery and it is a function of the current SoE . The original $SoE - \Delta SoE$ function is non-linear. However, a piecewise linear approximation is proposed in⁴⁰ to linearly approximate this relationship by:

$$soe(s,t) = \sum_{k=1}^{K-1} soe(s,t,k) \quad \forall s, \forall t \quad (20)$$

$$soe(s, t, k) \leq R(k + 1) - R(k) \quad \forall s, \forall t, \forall k \quad (21)$$

$$\Delta soe(s, t) = L_1 + \sum_{k=1}^{K-1} \frac{L(k + 1) - L(k)}{R(k + 1) - R(k)} soe(t - 1, s, k) \quad \forall s, \forall t \quad (22)$$

$$e^{ch,b}(s, t) \leq \frac{\Delta soe(s, t)}{\eta_b} \quad \forall s, \forall t \quad (23)$$

As can be seen the $SoE - \Delta SoE$ function is divided into $K - 1$ and in which each segment the original non-linear relationship is represented via a linear function. More details can be found in⁴⁰.

As can be concluded from this section, the system is formulated as a mixed integer linear problem (MILP). In order to solve the two-stage stochastic optimization, the scenarios for day-ahead PV energy generation are needed as the input.

4 | CASE STUDY AND SYSTEM PARAMETERS SELECTION

In this paper, the data set for local PV generation and building energy consumption of EnergyVille-1, a research center located in the city of Genk in Belgium, is used to evaluate the proposed method. A 368-kWp PV system, with 338 kVA AC rating, has been installed on the flat roof of the building. This building includes some energy metering devices to measure the real-time data related to the PV energy generation⁴¹, as well as local consumption. All the data are available with 15 minutes resolution, since April 2016.

The input parameters to scenario generation, as well as the optimization problem are described as follows:

1. PV system: The data from 9th April 2016 to 2nd February 2020 with the resolution of 15 minutes for the energy generation of the PV system installed in EnergyVille-1 have been considered in this paper. It should be noted that the measurement system has been installed on the AC side of the inverters; as a result there is no need to model the efficiency of inverters. The data set is splitted into three categories with the splitting range of 50%, 25%, and 25%, for training set, validation set and test set, respectively. Data from 1st March 2019 to 2nd February 2020 has been marked as test set and the rest of the data set has been divided into training set and validation set. Consequently, all the simulations in Section 5 have been done on the test set data.
2. The demand of the building: The electrical load profile of the building form 1st March 2019 to 2nd February 2020 with 15 minutes resolution has been downloaded from the local measurement system. It should be noted that the yearly energy demand of the building is much higher than the yearly local PV generation.
3. Electricity price: The day-ahead electricity market price from 1st March 2019 to 2nd February 2020 have been downloaded via the platform of the European Network of Transmission System Operators (entso)⁴². A 20% penalty in the price is assumed for real-time trading with the main grid, i.e., $C^{pur,rt}(t) = 1.2C^{da}(t)$ and $C^{sold,rt}(t) = 0.8C^{da}(t)$.
4. Battery: A 125 kWh battery is considered as the flexible asset of the building. All the parameters of the battery are selected based on the available information provided by⁴⁰. A nonlinear $SoE - \Delta SoE$ function is approximated by four piecewise linear sub-functions. The parameters of piecewise linear approximation function are shown in Figure 3. The x-axis and y-axis of Figure 3 represent the corresponding values for $R(k)$ and $L(k)$ in (22), respectively. η_b is set to 90% and the initial SoE is set to 10% of the capacity. It is noteworthy that at the end of horizon, the SoE reaches to its minimum value, since from prospective of optimization problem it does not make sense to store extra energy while it can be sold to the grid or consumed locally.

Since in this paper the proposed method is evaluated on the local real measured data set of EnergyVille-1, it is important to check the quality of the data (filling missing values and removing outliers) before starting simulations. Linear interpolation was applied to fill the missing values in each time series and the outliers were replaced by their logical boundary. In this paper, the simulations are done with 15 minutes resolution and over one day horizon ($T = 96$). All the simulations have been done in python on an Intel(R) Core i7-7500U CPU @2.70 GHz computer with installed memory 16 GB. In order to implement LASSO, the LASSO model from scikit-learn was used. The FFNN was implemented in Keras. The optimization framework has been modelled in PYOMO⁴³ while GUROBI 9.0.2⁴⁴ was used as the solver.

FIGURE 3 Parameters of the assumed piecewise linear approximation of $SoE - \Delta SoE$ function.

5 | SIMULATION RESULTS

In this section, the proposed method is applied to the available data set explained in Section 4. As mentioned before, this paper attempts to propose a scenario generation techniques, which is only based on historical PV measurement data. Moreover, the quality of generated scenarios is analyzed in the application of stochastic optimization for day-ahead schedule a battery. Therefore, in the following of this section, first we will show that a mixture of LASSO and empirical quantiles can lead to improve the accuracy of the probabilistic forecast. Then, by obtaining the distribution of the random variable, the scenarios are generated and the quality of them are analyzed in the stochastic optimization to manage energy in a PV-battery system to provide unbiased and stable solution.

5.1 | Predictive marginal distribution

In this sub-section, the simulation results regarding the predictive marginal distribution for PV energy generation are presented and discussed. First, the results of feature selection using LASSO technique is discussed. Second, the implementation of FFNN is explained in Section 5.1.2. Finally, the results of probabilistic forecast are discussed and evaluated in Section 5.1.3. It should be mentioned that the night-time data (from 7 in the evening to 7 in the morning) for PV energy generation is removed from the original data set, since there is almost no production during the night. Therefore, the forecast horizon includes 48 time step ($T = 48$).

5.1.1 | Feature selection

The first step to obtain the marginal predictive distribution is to select important features and determine the optimal lag, l in (1). In order to obtain a good generalization performance[‡], 10-fold cross validation was used to tune the parameter. Also, the training set itself was divided into two sub-sets: the first sub-set was used to train the model and the second sub-set was used as the validation set to select the optimal lag.

The real measurement data were normalized with respect to the nominal capacity of the inverter. Figure 4 illustrates the performance of LASSO over different lags for training set and validation set. The x-axis of Figure 4 represents different lags, from one day ago ($l = 1$) to three weeks ago ($l = 21$) while the y-axis represents NRMSE. As Figure 4 shows, the error of training set decreases by adding more lags. However, for validation set the error starts to decrease as well, until we reach to a certain point ($l = 13$) and after that the error over validation set starts to increase (overfitting issue). As the result, according to the NRMSE on validation set, $l = 13$ is selected as the optimal lag.

[‡]The machine learning model does not only learn to model the training data, but also should be generalized to new data, which has not seen before.

FIGURE 4 The performance of LASSO technique over training set and validation set for different lags in days. Lags started from one day ago (1) to three weeks ago (21).

It should be noted that in order to increase the training performance of LASSO, for each time step t , the historical data from 1 hour before to 1 hour after that particular time step were included. This results in 9 data points; from $t - 4$ to $t + 4$, for each lag. The simulation results for LASSO showed that for the optimal lag ($l = 13$), 47 out of 117 (13×9) coefficients (β s) ranked zero which were removed for the further steps.

5.1.2 | Point forecast

Here the optimal architecture for FFNN, which gives us the lowest NRMSE is discussed. The performance of our point forecast is discussed later in Section 6. The number of neurons for input and output layers should be equal to the dimension of our input and output matrix. Therefore, 70×48 and 48 neurons are needed in input layer and output layer, respectively. It is noted that 70 is the number of non-zero coefficients of LASSO and 48 is the number of time steps in one day after removing the night-time. Grid search is applied to tune the hyper-parameters, such as the number of hidden layer(s), the number of neuron in each hidden layer(s), as well as activation function. Different number of neurons are given as the input of grid search, from 0 to 300 with increase step 50. Also, different types of activation functions, including Sigmoid, Tanh, and Softplus function, are provided for grid search to search for the best combination of the input parameters. A FFNN including 2 hidden layers with 200 and 150 neurons in each layer and Sigmoid activation function results in the best performance with $\text{NRMSE} = 11.89\%$. It should be pointed out that the proposed method for PV forecast requires minimum amount of data. To compare the result of proposed method with other developed methods and see how accurate it is, in⁴⁵ the NRMSE reached to 10.14% for day-ahead PV forecaster, which was build based on historical weather data and historical PV measurements. In⁴ by using forecast values of solar irradiation as well as historical PV data, the performance of day-ahead PV forecaster was $\text{NRMSE} = 11.42\%$. Therefore, as can be seen, the accuracy of the proposed method in this paper is compatible with the other methods, while the advantage of proposed method is that it does not require any weather information.

5.1.3 | Predictive marginal distribution

The day-ahead predictive marginal distribution of PV energy generation for each time step is obtained via the quantiles of the relevant errors of the point forecast. Given the historical PV generation from Section 4, as well as historical point forecast of the system obtained from sub-section 5.1.2, the historical error for each time step and each day in the training set can be calculated. In order to calculate the quantiles of the error, for each targeted day and for each time step, only the historical errors of the selected important features from LASSO are considered. The quantile function of the errors are described by 99 quantiles with probability from 0.01 to 0.99 with step 0.01. Therefore, the predictive marginal distribution of PV generation can be obtained for each time step of the day-ahead.

FIGURE 5 Calibration diagram for all forecast horizons for proposed probabilistic forecast versus the benchmark method where the quantile function is obtained without using LASSO.

To evaluate the performance of probabilistic forecast, the calibration diagram is used. This diagram compares the similarity between the distribution of the real measurement (the x-axis) and the predicted distribution. The closer to diagonal, the more reliable results.

Figure 5 shows (average) calibration diagram for all the forecast horizon over the whole days in test set. In Figure 5, the solid black line belongs to the ideal case, the dot blue graph shows the results obtained by proposed method (the combination of LASSO and empirical quantiles of the error), while the gray dot graph illustrates the results of obtaining quantiles of the historical errors without selecting important features by LASSO. As can be seen, the proposed method in low and middle probability is well-calibrated. There is only a quite low deviation from the ideal case in the high probability. As Figure 5 shows the proposed method has a good performance on the test set without using any weather data (exogenous data). However, the results for non-parametric distribution without using LASSO ended up with a S-shaped curve and uncalibrated reliability diagram. In order to quantify the above comparison, two metrics are calculated for the proposed method and its benchmark: the maximum deviation from diagonal and the mean absolute error. The maximum deviation for proposed method is approximately 3.5%, while it is about 5.5% for the benchmark. Moreover, the mean absolute error is calculated 1.5% for the proposed method, and 2.4% for and the benchmark. In conclusion, the proposed method is able to improve the performance of probabilistic forecast.

Applying the above method gave us the predictive marginal distribution of PV energy generation over day-ahead. This indicates that for each time step of day-ahead, the individual predictive marginal distribution can be obtained ($\hat{F}_t(p_t)$).

5.2 | Scenario generation

In order to generate scenarios, the predictive joint distribution of the PV energy generation is needed. Gaussian copula was used in this paper to obtain multivariate distribution of our random variable. By transforming from original random variable space to multivariate Gaussian space, the empirical covariance matrix can be calculated.

The next step is to generate scenarios by sampling from the given multivariate distribution. Later, those generated samples are transformed to original random space. As previously mentioned, the quality of the generated scenarios can be evaluated by using two multivariate scoring metrics: ES and VS. As described in sub-section 2.6, these two metrics are proper and negatively oriented which means a lower score indicates a better performance. Figure 6 visualizes the results for ES and VS metrics over different number of scenarios, from 1 to 150. By looking at Figure 6, it can be seen that adding more number of scenarios could improve the both scoring metrics. This improvement is more noticeable for the lower number of scenarios (approximately less than 30 scenarios).

Generally, the higher number of scenarios means better representation of the random variable. On the other hand, the more scenarios are generated, the more computational time we will need to solve the stochastic problem at hand.

FIGURE 6 Evaluation of generated scenarios over different number of scenarios by using two multivariate scoring metrics reported in logarithm scale: Energy score (ES) on the left-hand side of y-axis; and Variogram-based score (VS) on the right-hand side y-axis. The values on the both y-axis are shown in logarithmic scale.

5.3 | Stochastic optimization

This sub-section presents the results of the two-stage stochastic optimization formulated in Section 3. First, we evaluate the stability and the bias of the obtained solution over the whole test set, which includes 339 days. In-sample and out-of-sample stability are analyzed to decide on the minimum number of scenarios. Second, in order to investigate the impact of the generated scenarios, the optimal value of different decision variables are discussed over one representative day. Moreover, to compare the results of the proposed method, five different benchmark cases are defined, where:

- Case1: It is assumed to know the exact day-ahead PV energy production.
- Case2: In this case the day-ahead PV generation is obtain via persistence method, which means $S = 1$ (in-sample analysis) while $N = 400$ (out-of-sample analysis).
- Case3: The point forecast values (obtained from LASSO+ FFNN) are used as the realization from day-ahead PV generation to schedule PV-battery system, which means $S = 1$ while $N = 400$.
- Case4: A deterministic Monte Carlo simulation is run. This means that in this case study we are facing with one-stage optimization (not two-stage) and the final results are the average cost over these 400 scenarios. In this case, the real-time electricity price (including markup) is assumed.
- Case5: This case is similar to the case4, except here the day-ahead electricity price is provided to run Monte Carlo simulation.

Figure 7a shows the average daily operational cost over different number of scenarios with in-sample and out-of-sample stability analysis over 339 days in our test set. In-sample analysis is focused on the results of the two-stage stochastic optimization when limited number of scenarios are used to represent the uncertainty. On the other hand, in out-of-sample analysis, large number of scenarios describe the random variable and the first-stage variables are fixed from in-sample analysis. Therefore, in-sample stability will be reached if the value of objective function does not change by increasing in the number of scenarios (x-axis in Figure 7a). Out-of-sample stability will be reached if the value of objective function obtained on a large number of scenarios ($N=400$ in this paper) does not change, while considering the first-stage variables fixed to their values from the stochastic optimization. As shown in Figure 7a, by adding more scenarios the gap between in-sample analysis and out-of-sample analysis decreases and in $S=25$, both in-sample and out-of-sample stability are reached. The results show a stable and unbiased solution with narrow width (approximately 2.4×10^{-2} percentage of the objective value) of the 95% confidence intervals around out-of-sample results (see Figure 7b). The computation time on the right-hand side of y-axis increased by adding more scenarios. Moreover, there is no (very small) difference between the value of objective function for in-sample and out-of-sample analysis,

(a) In-sample and out-of-sample stability analysis over 339 days in our test set. The results show a stable and unbiased solution. The computational time on the right-hand side of y-axis increases by adding more scenarios (S), while $N=400$.

(b) The 95% confidence intervals around out-of-sample results for different number of scenarios.

FIGURE 7 Quality analysis of the generated scenarios based on the proposed method.

TABLE 1 Average daily cost over different benchmark methods

Analysis	Case1	Case2	Case3	Case4	Case5	Proposed method
in-sample	261€	305.51€	306.73€	325.84€	304.08€	308.46€
out-of-sample	-	309.93€	309.73€	325.96€	304.22€	308.49€

which means that the solution is unbiased. The bias is around 0.004% (0.012€) of the value of the cost) at $S=25$, which is very small. It should be pointed out that for in-sample stability analysis each time we generate a new set of scenarios. However, the 400 scenarios for out-of-sample stability analysis are the same for each simulation result that corresponds to the same calendar day.

Table1 shows the results of benchmark methods together with the average daily cost of the proposed method ($S = 25$ and $N = 400$). The perfect information (Case1) benchmark can give an estimation on how we are far from this hypothetical benchmark. Moreover, as can be seen in Table1, Case2 and Case 3 end in biased solution. There is 1.43% (4.42€) bias between in-sample and out-of-sample analysis in Case2 (persistence method). Although, this bias drops to 0.98% (3€) in Case3, which shows the advantage of having point forecast; there is still a gap between the value of objective function for in-sample and out-of-sample analysis while the bias value for the proposed method is around 0.004% (0.012€). Moreover, in order to see the advantage of the two-stage stochastic PV-battery schedule, two deterministic Monte Carlo simulations are run over 25 and 400 generated scenarios in Case4 and Case5. As shown in Table1, in Case4 the difference between in-sample and out-of-sample analysis is around 0.04% (0.12€), which decreases compared to Case2 and Case3. However, in Case4 the average operational cost increases 5.4% (17.47€) compared to the proposed method, since the real time prices are considered. Therefore, by comparing the result of the proposed method with Case4, the advantage of two-two stage stochastic optimization with possibility of day-ahead decision making can be seen. Also, Case5 can give us an estimation on how far we are from this theoretical benchmark, where the scenarios are known in day-ahead time. As shown in Table1, the objective value of the proposed method is located between the results of Case4 and Case5. Therefore, the proposed method is not only able to reduce the bias of the solution, but also decrease the operational cost.

After analyzing the stability and the bias of the solution over the whole test set and showing its superiority compared to the benchmark methods, in order to see the optimal behavior of the battery as well as optimal transaction with the grid, the simulation result of one representative day (3rd June 2019, which was selected randomly) is shown in Figure 8. The results of point forecast, probabilistic forecast and scenario generation, together with the real PV production on 3rd June 2019 are shown in Figure 8a. For a better visualization only two quantiles of the probabilistic forecast with probability 0.1 (Q10) and 0.9 (Q90) are

(a) The results of random variable modeling, including point forecast, two predictive quantiles with probability of 10%(Q10) and 90%(Q90), and 10 generated scenarios.

(b) Input data for 3rd June 2019; total energy demand of the building on the left-hand side of y-axis; and the electricity price on the right-hand side y-axis.

(c) The day-ahead interaction with the main grid

(d) The state-of-energy (SoE) of the battery over 400 scenarios. Only the minimum and maximum values together with the quartiles are shown.

(e) Purchased energy in real time over 400 scenarios. Only the minimum and maximum values together with the quartiles are shown.

(f) Injected energy in real time over 400 scenarios. Only the minimum and maximum values together with the quartiles are shown.

FIGURE 8 Input data and simulation results for 3rd June 2019 as a representative day in the test set.

presented in Figure 8a. Also, for better visualisation only 10 scenarios are generated. Figure 8b shows the energy consumption of the building and day-ahead electricity price for buying and selling electricity during this representative day.

The optimal day-ahead interaction (first-stage variable) between the building and main grid is shown in Figure 8c. As can be seen, there is zero sold energy to the grid which means that optimization has decided to wait and see the realization of the PV energy generation. On the other hand, the day-ahead purchased energy is following the demand of the building except during peak load and high electricity price.

Figure 8d shows the optimal SoE of the battery over 400 number of scenarios. For better visualization, the quartiles of these 400 scenarios together with minimum and maximum values are shown in Figure 8d. As can be seen, in all scenarios the battery starts charging early in the morning when the price for electricity is low. The charging stops when the battery is full. In the morning, by increasing the price of the electricity and reaching the load of the building to its peak value, while the PV production has not reached to its maximum, in most of the scenarios the battery is discharged to supply the peak load while avoiding buying electricity. It can also be seen that in the afternoon by decreasing the load and electricity price and increasing the PV production, the battery starts charging. Later, the battery is discharged to supply the demand in the evening when there is no PV production and the electricity price is high. As the result, the battery is not only able to increase the self-consumption of the building, but also be able to reduce the operational cost of the building by avoiding buying electricity during the peak electricity price.

The real-time interaction of the building with the main grid is presented in the last two figures (8e, 8f). For better visualization, the quartiles of the 400 scenarios together with minimum and maximum values are shown. In all scenarios, the battery management system decides to buy electricity in the morning when the electricity price is at the lowest value and then there is no purchased energy during high price before noon. In the afternoon when the price decreases, in some scenarios (less than 50% of them), the building buys some amount of electricity from the grid and then it stops in the evening when the price is at its highest value. For real-time injection (see Figure 8f), in approximately 50% of the scenarios, the injection value is zero.

6 | DISCUSSION

In order to manage energy flow and available flexibility in a building including local PV installation, it is essential to model the uncertainty of the PV energy generation. In this paper, a data-based method was proposed to forecast the uncertainty of the PV energy generation in the form of possible scenarios. The proposed method was built based on historical real PV measured data which were collected locally in a smart building, office building, without using any exogenous data. The forecast of the uncertainty of the local PV system provides opportunity for the energy management system of the building to take decisions in uncertain environment which leads to more realistic and accurate results. The uncertainty of the PV system was forecast and shown as a set of scenarios, which are the discrete realization of our random variable. Afterwards, the energy management system of the building could use those generated scenarios to schedule the available flexible assets and to reduce the operational cost of electricity as well as decrease the peak load of the building.

The main advantages of the proposed method were that beside providing a stable and unbiased solution, there was no need for real and/or forecast of the weather data as well as PV forecast information provided by the external party, since those data, such as radiance and cloud cover forecasts or their measurements, are not available for free at the level of a building. The proposed method started to build point forecast only based on historical PV measurement data, which were measured locally. Then the predictive distribution was derived from point forecast to be translated to a scenario set afterwards. Moreover, as also discussed before, there was no need to have an assumption on the type of the distribution of the random variable. In fact, the distribution was obtained based on collected data, instead. However, the need for historical measured data for PV generation system makes difficult or impossible to apply the proposed method to the newly constructed buildings or the existing buildings without local (detailed) metering system. In those cases the historical data is not available and therefore it is not possible to train the model and forecast the distribution of random variable. Even though if the historical data is available, most of the users are not willing to share their detailed data due to the privacy issue. As a result, applying proposed method in those above cases will face a challenge. Non-disclosure agreement and providing a secure data transferring and analysis platform; as proposed in⁴⁶; can solve the privacy issue. Below, we discuss the challenge of applying proposed forecast uncertainty of PV energy generation in new buildings or existing buildings without any historical data.

To answer the question of how long after collecting data, the proposed method is able to reach to an accepted performance, the learning curve of FFNN is analyzed (see Figure 9). In general, learning curve shows the progress of learning (y-axis) over different size of the training set (x-axis). In this case, NRMSE is used as the evaluation metric to illustrate the losses of training.

FIGURE 9 Learning curve of the point forecast

As shown in Figure 9, the more data was provided to train FFNN, the lower error (or loss) results. Over approximately the first 100 days, the learning loss decreases drastically and after approximately providing one year of data, the NRMSE levels off. As the result, to be able to use the proposed method to forecast the uncertainty of generated PV system, at least 1 year of historical data of local PV energy generation is needed. It should be emphasized that to avoid overfitting one-third of the data was used as a validation set. Moreover, it is worth noting that this minimum value (1 year) on the size of data set was obtained based on the available data for the considered case study. However, this value can be different for another data set related to other location, since the PV generation and its predictability are highly influenced by weather conditions. Therefore, this approximation of one year data is valid for a system located in Belgium with high degree of cloud coverage.

If the required duration for data set is not available, the energy management system of the building can end up with two options to model the uncertainty of PV generation. The first one is to apply the persistence method where for each day the measured data of the previous day is used as the realization of PV energy generation for next day. In this situation, the uncertain behavior of PV system is ignored. The second option is to use the measured data of previous days as the possible scenarios to represent PV energy generation over next day and then these scenarios can be used as the input of stochastic optimization to schedule flexibility and manage energy and gradually move to the method proposed in this paper.

7 | CONCLUSION

In this paper, a two-stage stochastic optimization is proposed to manage energy in a PV-battery system, installed in a building, under PV uncertainty. Scenario set is used to describe the PV uncertainty for the stochastic optimization while this given scenario set affects the solution of stochastic optimization. As the result, it is important that the scenario set leads to a stable and unbiased solution in the stochastic optimization. A data-based approach is proposed to model the PV uncertainty and predict its distribution. As a pre-process, LASSO technique is applied to select important features and prepare data for the next steps. Afterwards, a FFNN is used to obtain point forecast for day-ahead PV energy generation. Then, the predictive marginal distribution, derived from point forecast; is calculated by a non-parametric approach. Finally the scenarios are generated from a multivariate distribution function obtained from Gaussian copula.

The proposed algorithm is applied on a data set related to a smart building and the simulation results show that the combination of LASSO and the quantiles of the historical forecast error can improve the accuracy of the non-parametric probabilistic error. Also, our case study shows that the proposed method to forecast day-ahead PV energy production and generate scenarios is able to provide a stable and unbiased solution for the stochastic optimization. The proposed method is only based on historical PV data and there is no need for any external data (weather data or PV forecast provided by external party). The proposed point forecast method reaches to NRMSE=11.89%, which is comparable with the results of other methods that are dependent on weather data. Also, low number of scenarios (25 scenarios) are needed to reach to in-sample and out-of-sample stability,

which result in decreasing the computation time. By comparing the results with benchmark cases, the proposed two-stage stochastic optimization cannot only reduce the average daily cost (5% cost reduction compared to the Monte Carlo simulation), but also decreases the bias of the solution, from 0.98% for point forecast benchmark to 0.004% for proposed method, while providing stability as well. The proposed method to obtain point forecast and generate scenario set is simple, which can be easily implemented in the building energy management system to manage PV-battery system.

The real implementation of the proposed algorithm to forecast day-ahead PV generation in Energy-Ville-1 building is the potential future work. This implementation can help energy management of the building to manage different flexible loads, while taking into account the uncertainty of PV system.

8 | ACKNOWLEDGMENTS

We would like to show our gratitude to Jesus Lago for assistance with the scenario generation part, to Jolien Despeghel for her comments and discussion about sizing the battery. This work was supported by the Internal Funds KU Leuven (C24/16/018). K. Bruninx is a post-doctoral research fellow of the Research Foundation - Flanders (FWO) at the University of Leuven and EnergyVille (mandate no. 12J3320N).

References

1. Agency IRE. Future of solar photovoltaic. <http://www.irena.org/-/media/Files/IRENA/Agency/Publication/2019> accessed: June 2020.
2. Mbuwir BV, Kaffash M, Deconinck G. Battery scheduling in a residential multi-carrier energy system using reinforcement learning. *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)* 2018: 1–6.
3. Kaffash M, Thoelen K, Deconinck G. Optimal Scheduling of a Residential Energy Storage Device in Multi-Carrier Energy Systems. *IEEE Benelux PELS/PES/IAS Young Researchers Symposium* 2018.
4. Kardakos EG, Alexiadis MC, Vagropoulos SI, Simoglou CK, Biskas PN, Bakirtzis AG. Application of time series and artificial neural network models in short-term forecasting of PV power generation. *2013 48th International Universities' Power Engineering Conference (UPEC)* 2013: 1–6.
5. Engels J, Claessens B, Deconinck G. Combined stochastic optimization of frequency control and self-consumption with a battery. *IEEE Transactions on Smart Grid* 2017; 10(2): 1971–1981.
6. Dutta S, Li Y, Venkataraman A, et al. Load and renewable energy forecasting for a microgrid using persistence technique. *Energy Procedia* 2017; 143: 617–622.
7. Weniger J, Bergner J, Quaschnig V. Integration of PV power and load forecasts into the operation of residential PV battery systems. *4th Solar integration workshop* 2014: 383–390.
8. Meer V. dD, Mouli GRC, Mouli GME, Elizondo LR, Bauer P. Energy management system with PV power forecast to optimally charge EVs at the workplace. *IEEE transactions on industrial informatics* 2016; 14(1): 311–320.
9. Nowotarski J, Weron R. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews* 2018; 81: 1548–1568.
10. Kaffash M, Deconinck G. Ensemble Machine Learning Forecaster for Day Ahead PV System Generation. *2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE)* 2019: 92–96.
11. Yona A, Senjyu T, Saber AY, Funabashi T, Sekine H, Kim CH. Application of neural network to 24-hour-ahead generating power forecasting for PV SYSTEM. *2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century* 2008: 1–6.

12. Niccolai A, Dolara A, Ogliari E. Hybrid PV Power Forecasting Methods: A Comparison of Different Approaches. *Energies* 2021; 14(2): 451.
13. Azuatalam D, Paridari K, Ma Y, Förstl M, Chapman AC, Verbič G. Energy management of small-scale PV-battery systems: A systematic review considering practical implementation, computational requirements, quality of input data and battery degradation. *Renewable and Sustainable Energy Reviews* 2019; 112: 555–570.
14. Pierro M, Bucci F, De Felice M, et al. Multi-Model Ensemble for day ahead prediction of photovoltaic power generation. *Solar Energy* 2016; 134: 132–146.
15. Bracale A, Carpinelli G, De Falco P, Rizzo R, Russo A. New advanced method and cost-based indices applied to probabilistic forecasting of photovoltaic generation. *Journal of Renewable and Sustainable Energy* 2016; 8(2): 023505.
16. Lorenz E, Heinemann D, Wickramarathne H, Beyer H, Bofinger S. Forecast of ensemble power production by grid-connected PV systems. *20th European PV Conference* 2007: 3–9.
17. Ma XY, Sun YZ, Fang HL. Scenario generation of wind power based on statistical uncertainty and variability. *IEEE Transactions on Sustainable Energy* 2013; 4(4): 894–904.
18. Bruninx K. Improved modeling of unit commitment decisions under uncertainty. *PhD thesis, KU Leuven* 2016.
19. El-Baz W, Tzscheutschler P, Wagner U. Day-ahead probabilistic PV generation forecast for buildings energy management systems. *Solar Energy* 2018; 171: 478–490.
20. Meer V, dDW, Widén J, Munkhammar J. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renewable and Sustainable Energy Reviews* 2018; 81: 1484–1512.
21. Michiorri A, Bossavy A, Kariniotakis G, Girard R. Impact of PV forecasts uncertainty in batteries management in microgrids. *2013 IEEE Grenoble Conference* 2013: 1–6.
22. Papaefthymiou G, Schavemaker P, Sluis V. dL, Kling W, Kurowicka D, Cooke R. Integration of stochastic generation in power systems. *International Journal of Electrical Power & Energy Systems* 2006; 28(9): 655–667.
23. Pinson P, Madsen H, Nielsen HA, Papaefthymiou G, Klöckl B. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology* 2009; 12(1): 51–62.
24. Su W, Wang J, Roh J. Stochastic energy scheduling in microgrids with intermittent renewable energy resources. *IEEE Transactions on Smart grid* 2013; 5(4): 1876–1883.
25. Yagli GM, Yang D, Srinivasan D. Reconciling solar forecasts: Probabilistic forecasting with homoscedastic Gaussian errors on a geographical hierarchy. *Solar Energy* 2020; 210: 59–67.
26. Golestaneh F, Gooi HB, Pinson P. Generation and evaluation of space–time trajectories of photovoltaic power. *Applied Energy* 2016; 176: 80–91.
27. Kaffash M, Deconinck G. Comparison of Statistical-Based and Data-Driven-Based Scenario Generation of PV Power for Stochastic Day-Ahead Battery Scheduling. *2020 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)* 2020: 730–734.
28. Kaut M, Wallace SW. Evaluation of scenario-generation methods for stochastic programming. *Pacific Journal of Optimization*; 2003.
29. Karevan Z, Suykens J. Spatio-temporal feature selection for black-box weather forecasting. *Proc. of the 24th european symposium on artificial neural networks, computational intelligence and machine learning* 2016: 611–616.
30. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 1996; 58(1): 267–288.

31. Shcherbakov MV, Brebels A, Shcherbakova NL, Tyukov AP, Janovsky TA, Kamaev VA. A survey of forecast error measures. *World Applied Sciences Journal* 2013; 24(24): 171–176.
32. He Y, Qin Y, Wang S, Wang X, Wang C. Electricity consumption probability density forecasting method based on LASSO-Quantile Regression Neural Network. *Applied energy* 2019; 233: 565–575.
33. Pinson P, Nielsen HA, Møller JK, Madsen H, Kariniotakis GN. Non-parametric probabilistic forecasts of wind power: required properties and evaluation. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology* 2007; 10(6): 497–516.
34. Bai L, Li F, Cui H, Jiang T, Sun H, Zhu J. Interval optimization based operating strategy for gas-electricity integrated energy systems considering demand response and wind uncertainty. *Applied energy* 2016; 167: 270–279.
35. Nelsen RB. *An introduction to copulas*. Springer Science & Business Media . 2007.
36. Angus JE. The probability integral transform and related results. *SIAM review* 1994; 36(4): 652–654.
37. Pinson P, Girard R. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy* 2012; 96: 12–20.
38. Woodruff DL, Deride J, Staid A, Watson JP, Slevogt G, Silva-Monroy C. Constructing probabilistic scenarios for wide-area solar power generation. *Solar Energy* 2018; 160: 153–167.
39. Bruninx K, Delarue E. Scenario reduction techniques and solution stability for stochastic unit commitment problems. *2016 IEEE International Energy Conference (ENERGYCON)* 2016: 1–7.
40. Pandžić H, Bobanac V. An accurate charging model of battery energy storage. *IEEE Transactions on Power Systems* 2018; 34(2): 1416–1426.
41. Yordanov G, Smolders F, Olaerts A, Verbeek G, Baert K, Driesen J. A 368-kWp Grid-connected PV System: Known and Hidden Losses. *Proceedings of the European PV Solar Energy Conference and Exhibition (EUPVSEC)* 2017: 25–29.
42. ENTSO . European Network of Transmission System Operators for Electricity. <http://transparency.entsoe.eu> accessed: 15 May 2020.
43. Hart WE, Laird CD, Watson JP, et al. *Pyomo-optimization modeling in python*. 67. Springer . 2017.
44. Gurobi Optimization I. Gurobi optimizer reference manual. <http://www.gurobi.com> 2018.
45. Monteiro C, Santos T, Fernandez-Jimenez LA, Ramirez-Rosado IJ, Terreros-Olarte MS. Short-term power forecasting model for photovoltaic plants based on historical similarity. *Energies* 2013; 6(5): 2624–2643.
46. Zhang XY, Kuenzel S. Differential Privacy for Deep Learning-based Online Energy Disaggregation System. *2020 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)* 2020: 904–908.

APPENDIX

A

Energy score (ES) is calculated by:

$$ES_t = \frac{1}{S} \sum_{s=1}^S \|p_t - \hat{p}_{s,t}\| - \frac{1}{2S^2} \sum_{s'=1}^S \sum_{s=1}^S \|\hat{p}_{s',t} - \hat{p}_{s,t}\| \quad (A1)$$

where in this paper ES is calculated for each time step of the day and is reported in average over forecast horizon T and total number of days D .

In this paper, the Variogram-based score (VS) is calculated by:

$$VS_t = \sum_{i,j=1}^T (\sqrt{|p_{t,i} - p_{t,j}|} - \frac{1}{S} \sum_{s=1}^S \sqrt{|\hat{p}_{s,t,i} - \hat{p}_{s,t,j}|})^2 \quad (\text{A2})$$

In simulation results, the average of VS is presented.