Short Term Load Forecasting in the Industry for Establishing Consumption Baselines: A French Case

José Blancarte, Mireille Batton-Hubert, Xavier Bay, Marie-Agnès Girard, and Anne Grau

Abstract The estimation of baseline electricity consumptions for energy efficiency and load management measures is an essential issue. When implementing real-time energy management platforms for Automatic Monitoring and Targeting (AMT) of energy consumption, baselines shall be calculated previously and must be adaptive to sudden changes. Short Term Load Forecasting (STLF) techniques can be a solution to determine a pertinent frame of reference. In this study, two different forecasting methods are implemented and assessed: a first method based on load curve clustering and a second one based on signal decomposition using Principal Component Analysis (PCA) and Multiple Linear Regression (MLR). Both methods were applied to three different sets of data corresponding to three different industrial sites from different sectors across France. For the evaluation of the methods, a specific criterion adapted to the context of energy management is proposed. The obtained results are satisfying for both of the proposed approaches but the clustering based method shows a better performance. Perspectives for exploring different forecasting methods for these applications are considered for future works, as well as their application to different load curves from diverse industrial sectors and equipments.

J. Blancarte (✉)
EDF R&D, Département Eco-efficacité et Procédés Industriels, Avenue des Renardières, 77818 Moret-Sur-Loing, France

Ecole Nationale Supérieure des Mines, UMR 6158, EMSE-Fayol, Saint-Etienne F-42023, France
e-mail: jose.blancarte@edf.fr; blancarte@emse.fr

M. Batton-Hubert • X. Bay • M.-A. Girard
Ecole Nationale Supérieure des Mines, UMR 6158, EMSE-Fayol, Saint-Etienne F-42023, France
e-mail: batton@emse.fr; bay@emse.fr; girard@emse.fr

A. Grau
EDF R&D, Département Eco-efficacité et Procédés Industriels, Avenue des Renardières, 77818 Moret-Sur-Loing, France
e-mail: anne.grau@edf.fr

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_1

1 Introduction

Establishing a baseline is the starting point to evaluate the potential as well as the results of different climate change mitigation related programs [31]. A baseline is the point of comparison to evaluate the behavior of different systems or procedures and allows to determine over or under performances. Thus, determining an energy consumption baseline is a key issue when implementing energy efficiency measures, deploying energy management programs, analyzing energy performance, and evaluating demand side management programs [8, 24, 31, 32].

When trying to determine if an industrial site or equipment is working under normal conditions, it is important to be able to compare their energy consumption with a "business as usual" forecasted one. This business as usual energy consumption is considered as the baseline scenario for comparison. This concept is particularly important in energy performance and efficiency contracts. The baseline allows the detection of abnormal consumption behaviors and/or overconsumption of equipments. Real-time monitoring of energy consumption helps an industrial site to optimize its energy consumption, reduce its costs, and adapt to changing electricity prices.

Energy efficiency has become a key parameter to be monitored by plant operators and managers aiming at optimizing their costs and reducing their energy losses [11]. Nowadays, most of the existing energy management platforms in the industry have a rather static nature, not adapting to real-time variability and having fixed thresholds, alarms and follow-up procedures. Energy management platforms should allow industrials to monitor their energy consumption and thus optimize their costs and detect abnormal behaviors on a real-time basis [16, 30].

Industrial sites are eager to implement energy efficiency recommendations. However, industry consumption may vary enormously from site to site and from sector to sector, and companies may deal with energy efficiency measures differently [1]. Added to this, there is a lack of relevant scientific literature for integrating energy performance in production management [4]. Baselines need to be consistently defined [31] and data analysis shall be as close as possible to standardized procedures in order to deploy energy management protocols faster and thus, reach as much industries as possible to increase the economical impacts due to energy efficiency [24].

Real-time energy consumption follow-up belongs to Automatic Monitoring and Targeting (AMT) techniques. AMT can be improved by the enhancement of the capabilities of the intrinsic data analysis methods used within an energy management platform. Adaptive methods for real-time energy consumption monitoring and analysis will lead to new methods of forecasting for establishing consumption baselines and thus, better energy consumption follow-up.

The main objective of this research study is to propose two different Short Term Load Forecasting (STLF) approaches for establishing a specific electricity consumption baseline on industrial data. The proposed techniques are applied for forecasting the power consumption of three different industrial sites from France,

from different sectors, at different moments of the day, and for short term forecasting horizons (2 h).

2 Materials and Data

For the purpose of this study, electricity consumption data was collected from three industrial sites from different regions in France (hereafter identified as sites A, B, and C). The three industrial sites belong to three different sectors and present different consumption patterns that are described below.

A big issue when implementing energy management programs is data availability. Generally speaking, energy consumption data at an industrial site level is monitored for billing and accounting purposes. This is not always the case with disaggregated data at workshop or equipment level, where metering instruments may be scarce. Other influential parameters are also not always monitored and thus not available on a first approach.

The only available monitored variable for the three sites is electricity consumption issued from historical billing data. The collected data is a 10 min interval load curve (each value being the average power consumption over a fixed 10 min interval). Each one of these intervals corresponds to each 10 min of the day from 12:00 am to 11:50 pm, which means 144 power consumption values for every available day. For the implementation of the different methods, the `R` software is used (N.B.: Due to confidentiality issues, orders of magnitude of the load curves have been omitted).

Site A

This particular site operates on an 8-h shift mainly from Monday to Friday, and in some occasions, on Saturdays. Not all weekdays present an operating electrical consumption activity, due to operational constraints of the site. Data is available for almost 2 years of electricity consumption. Figure 1 shows a 4-week interval of the load curve. Three main electricity consumption equipments are present at this site, which are turned on once the site is operating. Different equipment arrangements are operated as reflected in the load curve. For site A, 702 days of electricity consumption are available for analysis.

Site B

The second industrial site operates in a continuous 24 h cover, comprising three 8-h shifts. As it can be observed in Fig. 1, the consumption level might have big variations, since different workshops and equipments are engaged at different times

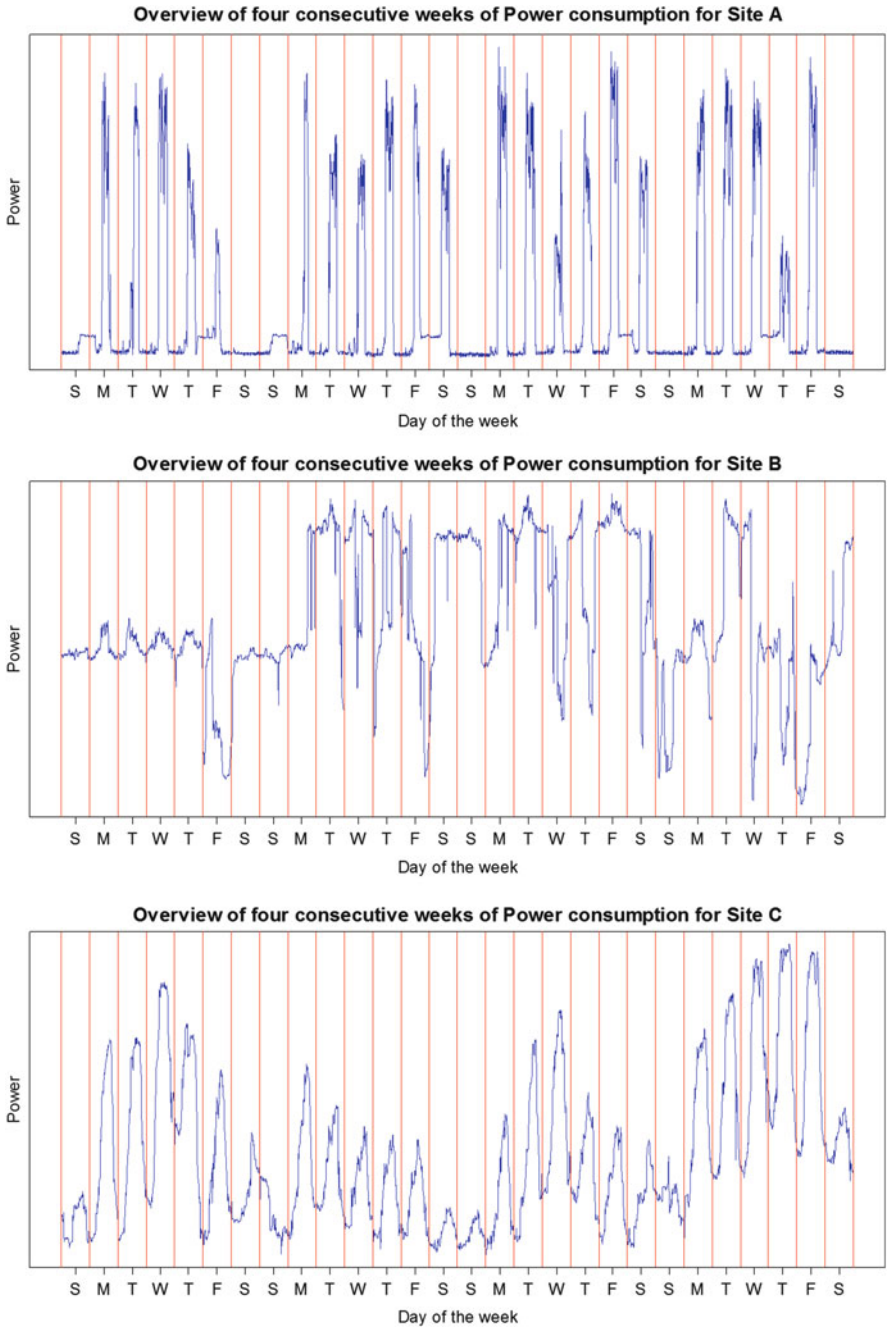

Fig. 1 Four weeks of electricity consumption for sites A, B and C

of the day. The load curve shape is significantly different than that of site A. For site B, 665 days of electricity consumption are available for analysis.

Site C

Site C is also a continuous 24 h cover industrial site. The consumption pattern is dependent on a daily activity as it is reflected in the load curve, shown in Fig. 1. During weekends, energy consumption is different than during weekdays. However, the electricity consumption base represents the biggest part of the consumed energy. Equipments keep consuming electricity during the night and during weekends in order to ensure certain operations at the site. For site C, 770 days of electricity consumption are available for analysis.

3 Forecasting Characteristics and Methods

Current electricity consumption forecasts are generally performed at a regional or national level, since their main interest is to ensure the efficient management of existing electrical power systems. National electricity loads have been at the core of electricity load forecasting for many years, and many techniques and methods have been proposed and assessed, as reviewed by many different authors [13, 15, 26, 27]. The different existing load forecasting methods can be classified into three main families: time-series analysis, multivariate analysis, and data-mining techniques [19]. However, when forecasting electricity consumption of industrial sites, the consumption may differ enormously in form, variability and influencing parameters for every single different site.

There is a lack of scientific literature focused on applying forecasting methods for establishing consumption baselines at lower consumption levels. Typical forecasting methods tend to be not well adapted when applied at an industrial site level. Seasonality, calendar events, and weather dependency are parameters usually taken into account when modeling a national electricity consumption curve [3]. However, due to the radically different nature of industrial sites, these parameters are inconsistent from site to site and may not be reflected in the consumption curve. Innovative approaches shall be followed to standardize the methods and have a larger reach and impact, as it was previously discussed.

When deploying energy management platforms in the industry, one of the main assumptions shall be that predictive models shall work with as little parameters as possible. As previously discussed, for industrial sites A, B, and C, the only available variable is historical electricity consumption. This section presents two different proposed forecasting techniques as well as the methods chosen to evaluate their relevance.

Table 1 lists all the symbols and parameters used in the text.

Table 1 List of symbols used in the text

Symbol	Definition
	Generic symbols
ξ	Gross energy deviations of the forecasted energy baseline with respect to the real energy consumption during the evaluation period (2 h)
i	Time period identifier
n	Time period at which a forecast is launched
t_n	Starting time of the forecast
P_i	Real power consumption at time period i
$\hat{P}_i$	Forecasted power consumption at time period i
θ	Forecast evaluation period (2 h)
N	Number of power consumption values during the evaluation periods (12)
TV_n	Truncated test vector up to the nth interval.
p	Dimension size of the individuals (144 power consumption points)
v	Number of intervals used to construct the adjustment factor
FAJ_v	Adjustment factor using v intervals
	Method 1
M_1,M_2	Dimensions of the SOM grid
m	Number of neurons
k	Identifier of the neuron
Cl_k	Coordinate vectors of the different neurons
$Cl_{k,h}$	Coordinate of the hth element of the kth neuron
$Cl_{k,n}$	Coordinates vector of the truncated cluster vectors up to the nth interval.
WN	Reference vector corresponding to the winning neuron (Also known as the BMU)
X_{tr}	Vector corresponding to the chosen element from the training data
λ	Number of iterations of the SOM algorithm
s	Current iteration step of the SOM algorithm
α	Learning rate of the SOM algorithm
σ	Radius of the neighborhood of the SOM algorithm
D_k	Distance of the updated node to the winning unit
Cl_W	Winning reference vector
	Method 2
Λ	Eigenvectors matrix
U	Eigenvalues matrix of the principal components
A	Covariance matrix
j	Number of principal components explaining 90 % of the data variability
q	Principal component identifier
U_q	Coordinates of the qth principal component
$\hat{C}_q$	MLR coefficients for the qth principal component
$\hat{\varepsilon}$	MLR disturbance coefficient
Un_q	Coordinates vector of the truncated principal component up to the nth interval

3.1 Forecasting Characteristics

For constructing the different models, every day is considered as an independent element (a vector) composed of 144 values of power consumption. Days can be considered to be independent for practical purposes: forecasts are performed intra-day and consumption cycles present daily patterns in most of the cases, corresponding to different consumption modes. Simple data splitting [22] is used for model validation. Eighty-five percent of the data (85 % of the available days for analysis) is used as the training dataset. The remaining 15 % (test dataset) is used to test the models and compare the performance of both methods. Data sampling of the days is performed randomly on a stratified manner at two levels, in order to obtain a distribution of different seasonal variabilities related to time parameters: day of the week and month of the year.

In order to test the different methods for power consumption forecasting at the site level and at different moments of the day, different parameters and characteristics for the forecasts have to be defined. For each test day, the baseline load is estimated at each hour from 9:00 am to 5:00 pm for site A and from 9:00 am to 9:00 pm for sites B and C. In order to evaluate the performance of the forecasting methods, the forecasting periods are fixed to be the following 2 h (called forecast evaluation period, identified by θ, composed of 12 power consumption intervals), considered as short term load forecasting (STLF). In short the different methods will forecast the power and energy consumption from a specific time-step (called t_n, which will be varied from 9:00 am to 5:00 pm or 9:00 pm, depending on the site) for a specific number of time intervals (called N, which has been defined as 12) that corresponds to 2 h.

In total, for site A, 882 simulations will be performed (98 test days, 9 simulations per day from 9:00 am to 5:00 pm), 1,170 for site B (90 test days, 13 simulations per day from 9:00 am to 9:00 pm), and 1,339 for site C (103 test days, 13 simulations per day, similar as for site B). The simulation results will be compared according to the performance indicator defined further on.

3.2 Proposed Forecasting Methods

If the objective is to analyze as many sites as possible, methods shall be easy to deploy and should not require much human input or expertise. Also, they shall demand low calculation times in order to easily standardize the procedures. To overcome these problems, two different approaches for electricity consumption forecasting are proposed, based on the nature of the examined data:

- A first method using load curves clustering in order to detect consumption patterns that will be used as electricity consumption forecasts.

- A second method based on signal decomposition in order to detect the variability of the daily behavior of the curves, using the eigenvectors issued from a Principal Component Analysis (PCA) of the training dataset.

3.2.1 Method 1: Electricity Consumption Forecasting by Pattern Recognition Using Load Curve Clustering

Electrical load curve clustering has attracted much attention in recent years for its application in client profiling and electricity pricing [6, 10, 20]. The capacity of clustering techniques for handling large amounts of time-series data has been assessed in the past [23]. Diverse clustering techniques have been used in the past, as reviewed by Chicco in [5]. From the different assessed clustering techniques, K-means and Self-organizing Maps (SOM) are the best performing ones. SOM is not a direct clustering method, as explained in [6], however, it produces a visually understandable projection of the original data into a map. In this study, SOM has been chosen as the clustering technique due to its prior application for forecasting purposes, as proposed by different authors [7, 18, 25]. Nevertheless, these previous work were focused in forecasting national electricity demand.

SOM [17, 23, 28] is an unsupervised neural network that projects a p-dimensional input dataset onto a reduced dimension space (one or two-dimensional in most cases). SOM is composed of a predefined grid of $M1 \times M2$ elements called neurons (m number of neurons). Each neuron (identified by k) is also p-dimensional. Neurons have to be initialized, this means, the p-dimensions of the k neurons have to be previously defined by a reference vector Cl_k, as in expression (1), where $1 \leq k \leq m$, and $P_{k,i}$ is the value of power consumption for element i of neuron k, where $1 \leq i \leq p$. Initialization of the SOM algorithm can be done in different manners (randomly or data analysis based initialization) as described in [2].

$$Cl_k = [P_{k,1}, P_{k,2}, \ldots P_{k,p}] \tag{1}$$

All neurons are associated to neighboring neurons of the map by a neighborhood relation, which determines the "area" of influence within the defined space. Neurons are calculated through a competitive algorithm that recalculates the weights of the winning neuron and the weights of its neighboring neurons proportionally inverse to their distance. The neighborhood size will be reduced at each iteration during the map training process, starting with almost the full topography and ending in single neurons.

Once all Cl_k reference vectors have been initialized, SOM training starts. The algorithm will be run a predefined number of iterations, represented by λ. At each iteration (represented by $s \leq \lambda$), an input vector X_{tr} (as described in formula (2)) issued from the training data set is chosen randomly, where tr goes from 1 to the number of individuals in the training dataset, and $P_{tr,i}$ is the value of power consumption for element i of the tr individual and where $1 \leq i \leq p$. Euclidean distances between the chosen X_{tr} and all the Cl_k vectors are calculated. The closest

reference vector is known as the winning neuron (*WN*) or best matching unit (BMU), as in expression (3).

$$X_{tr} = [P_{tr,1}, P_{tr,2}, \ldots, P_{tr,p}] \tag{2}$$

$$WN = argmin \left\{ \sqrt{\sum_{i=1}^{i=p} (X_{tr,i} - Cl_{k,i})^2)} \right\}; 1 \leq k \leq m \tag{3}$$

The coordinates of *WN* and its neighboring neurons are adjusted then towards the coordinates of the input vector X_{tr}, as in expression (4), where $\alpha(s)$ is the learning rate which decays with each iteration, and $\theta(s)$ is the neighborhood function, represented in expression (5). The radius $\sigma(s)$ is also updated at every iteration, shrinking over time. D_k is the distance of the updated node to *WN* (the winning neuron).

$$Cl_k(s+1) = Cl_k(s) + \alpha(s)\theta(s) \cdot (X_{tr}(s) - Cl_k(s)) \tag{4}$$

$$\theta(s) = exp(-\frac{D_k^2}{2\sigma^2(s)}) \tag{5}$$

The proposed methodology based on pattern recognition using SOMs is described below and divided into three steps:

1. Once the data splitting has been performed as described previously, the training dataset will be used to construct the reference vectors.

The SOM algorithm is launched considering the daily load curves as individuals for analysis ($p = 144$). As defined previously, the SOM algorithm needs a number of clusters (m) before its initialization. Tsekouras et al. [29] have determined that for large electricity customers 8–12 clusters are necessary for a satisfactory description of the daily load curves. Different numbers of clusters will be tested in order to determine a good description of the different possible load curves. The algorithm is performed on non-reduced data, as suggested in [10]. For the purpose of this study, in order to converge to the same solution, a linear initialization is used. A rectangular configuration of the neighborhood is chosen due to its visualization effectiveness. The chosen number of iterations is $\lambda = 100$. The chosen learning rate is a linear function from 0.05 to 0.01 over the 100 iterations for which it was found that the algorithm converges rapidly. The neighborhood radius is varied from a value of two thirds of all unit to unit distances to its negative value, linearly through the different iterations. Once the neighborhood gets smaller than one individual, only the *WN* reference vector is changed.

The resulting Cl_k reference vectors of each neuron are then kept and assigned to the neuron according to the described SOM algorithm. Every cluster is then represented by a vector composed of 144 variables identified as Cl_k, the identifier of the cluster. $Cl_{k,h}$ contains the value of the hth variable of the kth neuron. Every

variable represents a specific power consumption point of the day, as defined previously.

2. Once the SOM algorithm has been performed, the Cl_k centroid vectors will be used. At the time (t_n) a forecast is simulated for a chosen individual of the test dataset, the test element is truncated to a vector (identified as TV_n) composed of n elements as shown in expression (6) (where $n \leq p$)

$$TV_n = [P_{n,1}, P_{n,2}, \ldots, P_{n,n}] \tag{6}$$

The different Cl_k vectors will be then truncated up to the nth element and called $Cl_{k,n}$. Euclidean distances will then be calculated for the TV_n vector to the different $Cl_{k,n}$ vectors. The vector corresponding to the minimum distance is then considered the winning vector, identified by Cl_W as in expression (7).

$$Cl_W = argmin \left\{ \sqrt{\sum_{i=1}^{i=n} (Cl_{k,i} - TV_{n,i})^2)} \right\}; 1 \leq k \leq m \tag{7}$$

3. The coordinates of the cluster Cl_W corresponding to that vector will be proposed as the forecast for the following consumption points. The forecasted power consumption points $\hat{P}_i$ will correspond to those the elements with the same index i of the closest Cl_W vector, represented by $Cl_{W,i}$ as expressed in formula (8).

$$\hat{P}_i = Cl_{W,i} \tag{8}$$

3.2.2 Method 2: Electricity Consumption Forecasting by Signal Decomposition Using Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate analysis technique used in many different areas for analyzing large sets of data [9, 14]. The pertinence of PCA coupled with other techniques for forecasting energy consumption has been assessed by some authors [21, 27]. However, for different applications, the PCA is used as a tool among others to produce a specific predictor, adapted to the nature of the data.

It can be assumed that the electricity consumption at the site level is a function composed of different signals. The changes and variability in electricity consumption can be explained by different variables. The PCA allows to obtain the eigenvalues (matrix Λ) that explain most of the variability of the data and the eigenvectors (matrix U) of the principal components which are obtained by the decomposition of covariance matrix A in ${}^tU\Lambda U$ that are uncorrelated to each other.

For the purpose of this study, the PCA is performed on the training data set (on non-reduced data). The coordinates of the j first eigenvectors explaining 90 % of the variability are preserved. These coordinates have in fact a meaning according to a specific power consumption interval of the day, since the reduced variables are

in fact power consumption time intervals. U_q are the coordinates of the different j eigenvectors, where $1 \leq q \leq j$, as expressed in (9).

$$U_q = [P_{q,1}, P_{q,2}, \ldots, P_{q,144}] \tag{9}$$

The preserved principal components are used to build a predictor based on their linear combination in order to predict the variability of the data. This is done by using a Multiple Linear Regression (MLR). At the time t_n a prediction is launched for a chosen element of the test dataset, the different U_q vectors are then truncated up to the point n of the forecast and represented by Un_q, as expressed in formula (10).

$$Un_q = [P_{q,1}, P_{q,2}, \ldots, P_{q,n}] \tag{10}$$

A multiple linear regression is used in order to find the coefficients for the principal components, for which the combination of these components fit the data of the chosen element of the test dataset. For this purpose, an ordinary least squares model is used. The vector TV_n is defined as a function ($P(i)$) determining the power consumption value at timestep i. Along with the different j truncated Un_q eigenvectors, $P(i)$ is used to fit a linear model as in expression (11). The coefficients $\hat{C}_q$ and the intercept term $\hat{\varepsilon}$ are obtained through the MLR.

$$P(i) = \sum_{q=1}^{j} \hat{C}_q Un_q(i) + \hat{\varepsilon} \tag{11}$$

These coefficients and the eigenvectors coordinates U_q are used for predicting the power consumption of the site for the rest of the day for every consumption point $\hat{P}_i$, as in expression (12).

$$\hat{P}_i = \sum_{q=1}^{j} \hat{C}_q U_{q,i} + \hat{\varepsilon} \tag{12}$$

3.3 Adjustment Factor

An adjustment factor can be used to improve the forecasts of different techniques. Method 1 is based on pattern recognition, and the proposed forecast is a typical energy consumption mode. However, even if the a pattern has been correctly recognized, the forecast may under or over estimate the actual consumption level corresponding to a specific day. An adjustment factor may deal with this problem by adjusting the forecast to the correct level of energy consumption. The adjustment factor deals as well with the issue of trends, in case they exist. Method 2 forecasts

are issued from a Multiple Linear Regression and thus, the different coefficients fit a model to the actual consumption level and no adjustment factor is needed.

Different forms of adjustment factors exist, but the most important ones can be classified in two different categories for univariate methods: scalar and additive, as described by different authors [8, 12]. Since only electricity consumption information is available for the concerned industrial sites, weather or other related adjustment factors will not be considered in this study.

To deal with the mentioned issues, a scalar adjustment factor is used to improve the forecasts of Method 1. The proposed adjustment factor is calculated as in expression (13), and corresponds to the average of the ratios of the v previous real power consumption values to predicted ones. P_i represents the real power consumption at time interval i, $\hat{P}_i$ is the forecasted power consumption at time interval i, and v the number of intervals used to construct the adjustment factor. The chosen number of intervals (v) for the adjustment factor is one, since better results are obtained in terms of the chosen performance indicator and since calculation times are reduced.

$$FAJ_v = [\frac{P_{i-1}}{\hat{P}_{i-1}} + \cdots + \frac{P_{i-v}}{\hat{P}_{i-v}}] \times \frac{1}{v} \tag{13}$$

3.4 Performance Indicator

The main indicators used in literature to evaluate the performance of forecasting methods are the Mean Absolute Percentage Error (MAPE) and the Mean Squared Error (MSE). These parameters are adequate when evaluating the resemblance of a forecast compared to a real curve. These indicators are adapted to situations where the goal is to optimize the use of production means to meet an electricity demand, or residual demand curves calculation in competitive electricity markets. This is not the case when evaluating the forecasting performance in an industrial site for energy efficiency purposes. Most energy efficiency programs have an economic constraint and are rewarded or penalized economically if objectives are met or not [31]. For this reason, a specific performance criterion is proposed and used which is easily transformed into an economic indicator.

This criterion is directly linked to the site's global energy consumption. It is based on the difference, in energy (kWh), between forecasted energy consumptions issued from the models and real energy consumptions issued from the data. The indicator is based on gross energy differences (hereafter referred as Gross Energy Deviation, GED) through the time period of the forecast (2 h), and represented by symbol ξ. GED and its distribution will be used to evaluate the relevance of each method. Expression (14) formalizes the way of calculating these deviations, where P_i is the actual power consumption at time-step i, and $\hat{P}_i$ is the forecasted power consumption at that same time-step. θ and N were previously defined.

$$\xi = \left[\sum_{i=t_n}^{t_n+N} P_i - \sum_{i=t_n}^{t_n+N} \hat{P}_i \right] \times \frac{\theta}{N} \tag{14}$$

GED allows to evaluate the distribution of the forecasts in terms of how much is the forecast above or below an energy threshold which is the real energy consumption during that time period. This approach is useful to set operational parameters and thresholds in the industry, and to easily translate them into an economic indicator.

4 Results and Discussion

The results for each of the implemented methods are described below. A focus is made on the evaluation of the performance of both forecasting methods presented above, according to the defined criteria.

4.1 Method 1: Electricity Consumption Forecasting Using Self-Organizing Maps

For the implementation of Method 1, different tests were carried out varying the neurons number from 8 to 12 (as in [29]), selecting the lowest number of neurons for which the GED distribution is does not vary greatly if another neuron is proposed. Twelve neurons were selected for site A, 9 for site B, and 12 for site C. The graphical representation of the different reference vectors for the sites can be seen in Fig. 2. The different identified patterns correspond to the different typical consumption modes of the sites. These typical load curves are used for forecasting as explained previously.

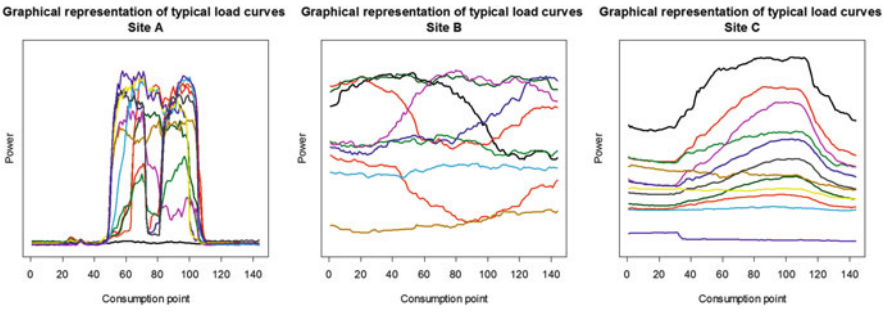

Fig. 2 Resulting curves after applying the SOM algorithm to the test dataset for the three sites

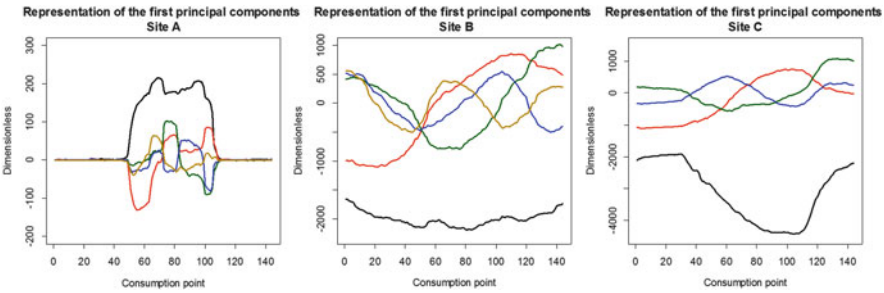

Fig. 3 Graphical representation of the main principal components coordinates for the three sites

4.2 Method 2: Electricity Consumption Forecasting Using Principal Component Analysis

For the implementation of Method 2, in order to explain 90 % of the variability of the data, the first 5 principal components are selected for sites A and B, and 4 principal components are selected for site C. Adding more principal components or increasing the 90 % threshold would increase calculation times, which is to be taken into account when monitoring energy consumption in real time. The footprint of the different components for the different sites can be seen in Fig. 3. These principal components are the ones used to run the MLR that will determine the coefficients for the forecasting models.

4.3 Results by Site

Results obtained using both methods for each of the studied sites are presented below.

Site A

The distribution of the different obtained GED for site A with Method 1 is presented in Fig. 4. The "y" axis represents the gross energy deviation and the "x" axis is the energy that was actually consumed during that period, in order to relativize the error of the forecast in terms of energy. Points outside the dashed lines are above a 50 % GED threshold, and points outside the solid lines are above a 10 % threshold. In order to evaluate the performance of the methods at differents times of the day, different hours were grouped into four different time-spans: from 9:00 am to 11:00 am (morning), identified by the solid green squares; from 12:00 pm to 2:00 pm (noon), identified by the solid pink circles; from 3:00 pm to 5:00 pm (early

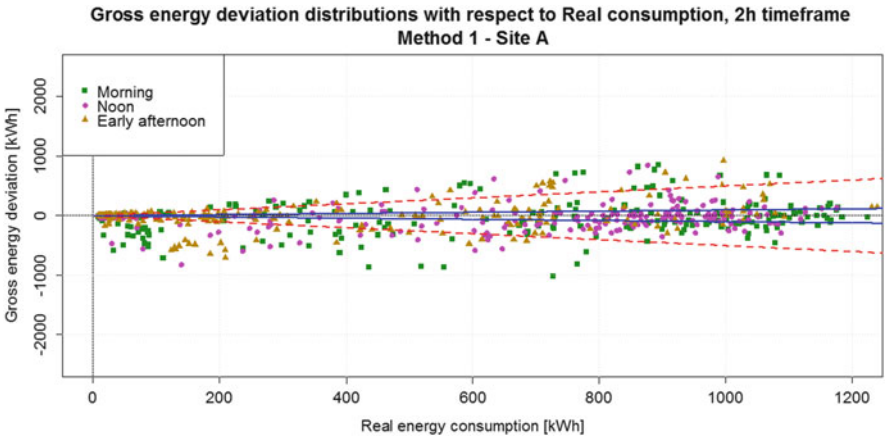

Fig. 4 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 1 for site A

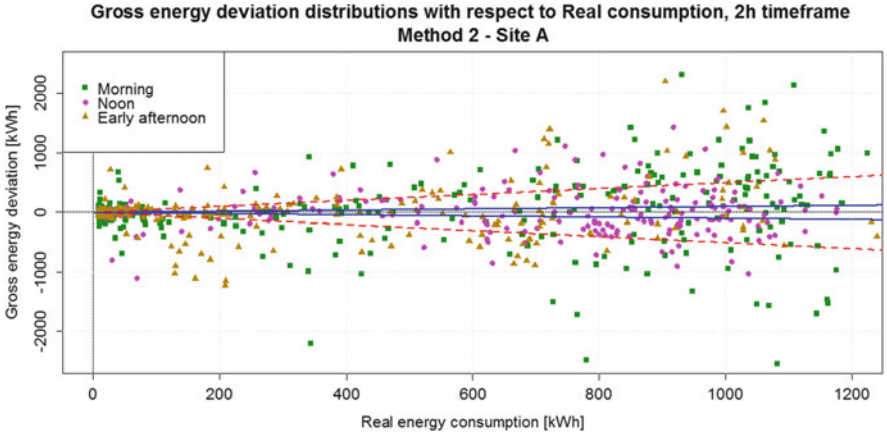

Fig. 5 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 2 for site A

afternoon), identified by the solid yellow triangles, and from 6:00 pm to 9:00 pm (late afternoon), identified by black crosses.

For Method 1, 226 simulation points are inside the solid lines and 642 are between the dashed lines. The total simulation points for this site are 882. Figure 5 represents the GED distributions for site A using Method 2. Solid lines and dashed lines represent the same thresholds as in Fig. 4. For Method 2, only 104 simulation points are inside the solid lines, while 410 are between the dashed lines.

Looking at the dispersion of the points and the number of them outside of the defined thresholds, of Figs. 4 and 5, Method 1 clearly outperforms Method 2 for this particular industrial site. Regarding the distribution of the different timespans,

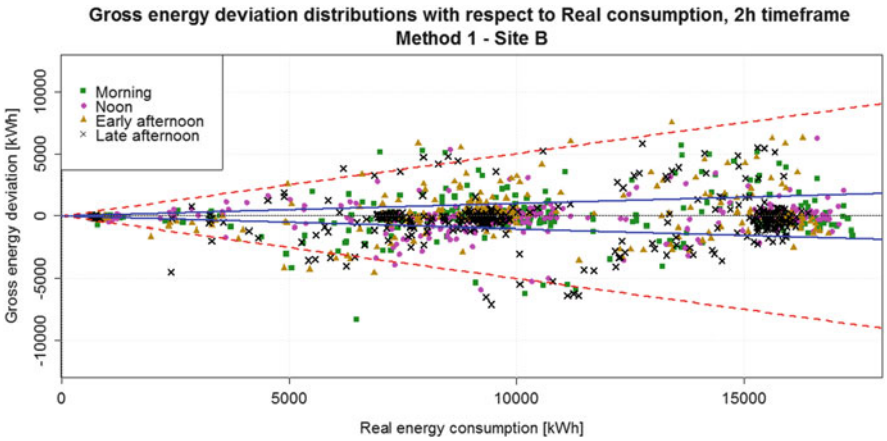

Fig. 6 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 1 for site B

besides a slightly wider distribution for the morning period, no significant difference can be observed for the different hours of the day.

Site B

Figure 6 shows the GED distribution for site B using Method 1 as in Fig. 4. For this site, 767 simulation points are inside the solid lines and 1,123 are between the dashed lines. The total simulation points for this site are 1,170.

Figure 7 represents the GED distributions for site B using Method 2. Solid lines and dashed lines represent the same thresholds as in previous figures. For Method 2, 467 simulation points are inside the solid lines, and 1,001 are between the dashed lines.

For this industrial site, Method 1 also outperforms Method 2. As for the distribution regarding the different timespans, no significant difference can be observed to conclude a strong influence of the hours of the day for both methods.

Site C

Figure 8 shows the GED distribution for site C using Method 1 as in Fig. 6. For this site, 1,146 simulation points are inside the solid lines, which represent less than 10 % in error, and 1,334 are between the dashed lines that represent less than 50 % in energy error. The total simulation points for this site are 1,339. It is important to notice that only five points are outside the dashed line boundaries in this particular case.

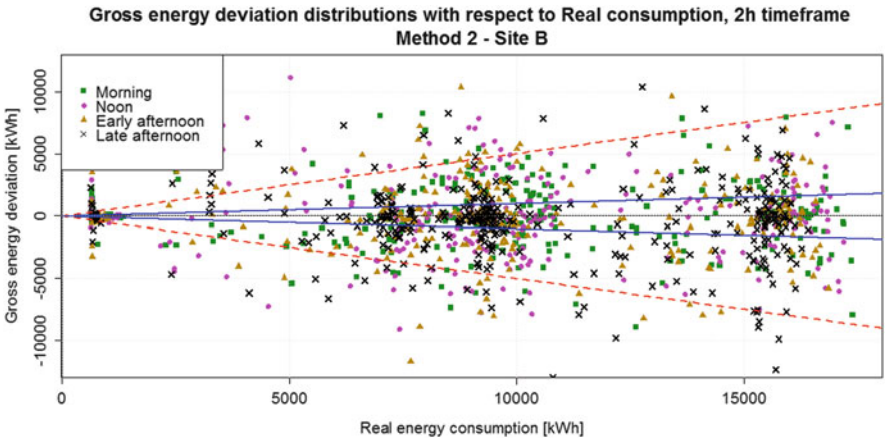

Fig. 7 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 2 for site B

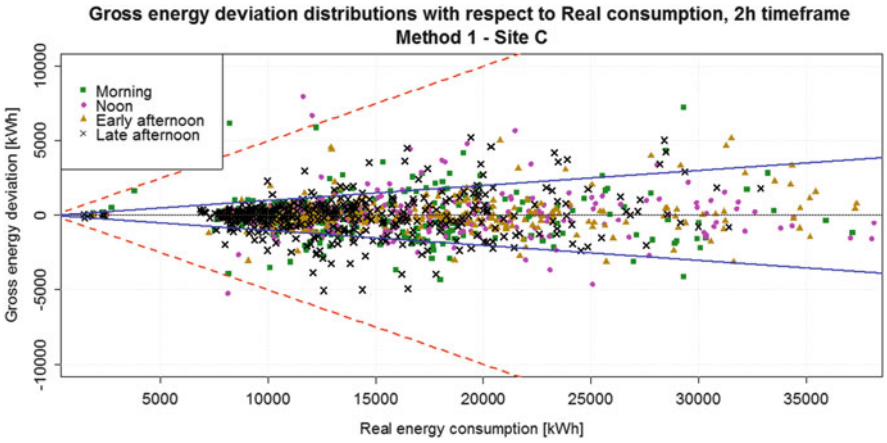

Fig. 8 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 1 for site C

Figure 9 represents the GED distributions for site C using Method 2. Solid lines and dashed lines represent the same thresholds as in previous figures. For Method 2, 884 simulation points are inside the solid lines, and 1,319 are below the 50 % threshold represented by the dashed lines.

Even though results can be considered satisfactory for Method 2 applied to industrial site C, Method 1 still shows better performances. As well as for sites A and B, the hour of the day does not seem to influence greatly the performance of the methods, since the GED distributions are evenly distributed for all of the timespans.

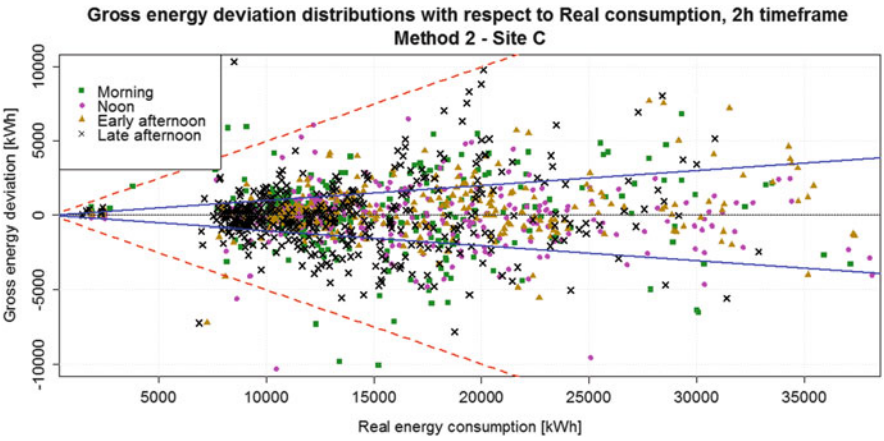

Fig. 9 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 2 for site C

5 Conclusions and Perspectives

Two different methods for establishing short-term electricity consumption baselines were proposed and assessed. From the obtained results, Method 1 outperforms Method 2 when forecasting the short term electricity consumption for the three presented industrial sites, according to the chosen performance indicator. Added to this, the hour of the day does not significantly influence the performance of the methods.

Subsequent works will focus on specific industrial equipments that are installed at the industrial sites and that contribute to most of their power consumption. The aggregation of industrial equipments allows a more flexible and adaptable energy consumption follow-up, since information can be lost at the industrial site level. In order to ensure the validity and repeatability of the obtained results for their generalisation, future research works will focus on the construction of a bootstrapping procedure.

Perspectives to improve the forecasting potential for Method 2, could be the integration of weighing factors for the coefficients and studying the errors obtained for the different forecasts at different times of the day.

Model combination could be a clue to improve the performance of the forecasts, since it could integrate different approaches (such as form recognition and Bayesian inference) in order to overcome the deficiencies of the different methods.

It is important to point out that due to the variability of the data, the differences from site to site and from sector to sector, standardizing the methods to build energy consumption baselines can be a hard task. The use of additional variables shall be considered when possible, which will make the methods more adaptable. Univariate

methods could rapidly reach a limit of performance. The main problem which may persist will be data availability.

Energy management can be improved by the utilization of different methods to calculate energy consumption baselines for the diverse energy management applications. Performing bottom-up approaches provides more precise information and makes energy consumption flexibility fast and reactive.

References

1. Alhourani, F., & Saxena, U. (2009). Factors affecting the implementation rates of energy and productivity recommendations in small and medium sized companies. *Journal of Manufacturing Systems, 28*(1), 41–45. doi:10.1016/j.jmsy.2009.04.001.
2. Attik, M., Bougrain, L., & Alexandre, F. (2005). Self-organizing map initialization. In *Artificial neural networks: biological inspirations – ICANN 2005*, Warsaw (pp. 357–362).
3. Bunn, D. W., & Farmer, E. D. (1985). *Comparative models for electrical load forecasting*. Chichester/New York: Wiley.
4. Bunse, K., Vodicka, M., Schönsleben, P., Brülhart, M., & Ernst, F. O. (2011). Integrating energy efficiency performance in production management – Gap analysis between industrial needs and scientific literature. *Journal of Cleaner Production, 19*(6–7), 667–679. doi:10.1016/j.jclepro.2010.11.011.
5. Chicco, G. (2012). Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy, 42*(1), 68–80. doi:10.1016/j.energy.2011.12.031.
6. Chicco, G., Napoli, R., & Piglione, F. (2006). Comparisons among clustering techniques for electricity customer classification. *IEEE Transactions on Power Systems, 21*(2), 933–940. doi:10.1109/TPWRS.2006.873122.
7. Cottrell, M. (2003). Some other applications of the SOM algorithm: How to use the Kohonen algorithm for forecasting. In *Invited lecture at the international work-conference on artificial neural networks, IWANN 2003:* Maó, Menorca, Spain.
8. Coughlin, K., Piette, M. A., Goldman, C., & Kiliccote, S. (2009). Statistical analysis of baseline load models for non-residential buildings. *Energy and Buildings, 41*(4), 374–381.
9. Daultrey, S. (1976). *Principal components analysis*. Norwich: Geo Abstracts.
10. Fidalgo, J. N., Matos, M. A., & Ribeiro, L. (2012). A new clustering algorithm for load profiling based on billing data. *Electric Power Systems Research, 82*(1), 27–33. doi:10.1016/j.epsr.2011.08.016.
11. Giacone, E., & Manc, S. (2012). Energy efficiency measurement in industrial processes. *Energy, 38*(1), 331–345. doi:10.1016/j.energy.2011.11.054.
12. Goldberg, M. L., & Kennedy Agnew, G. (2003). *Protocol development for demand response calculation: Findings and recommendations* (Technical report). KEMA-Xenergy.
13. Hahn, H., Meyer-Nieberg, S., & Pickl, S. (2009). Electric load forecasting methods: Tools for decision making. *European Journal of Operational Research, 199*(3), 902–907
14. Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary: Sas Institute.
15. Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems, 16*(1), 44–55.
16. Hu, S., Liu, F., He, Y., & Hu, T. (2012). An on-line approach for energy efficiency monitoring of machine tools. *Journal of Cleaner Production, 27*, 133–140. doi:10.1016/j.jclepro.2012.01.013.
17. Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78*(9), 1464–1480. doi:10.1109/5.58325.

18. Lendasse, A., Lee, J., Wertz, V., & Verleysen, M. (2002). Forecasting electricity consumption using nonlinear projection and self-organizing maps. *Neurocomputing, 48*(1), 299–311.
19. Li, D. C., Chang, C. J., Chen, C. C., & Chen, W. C. (2012). Forecasting short-term electricity consumption using the adaptive grey-based approach – An Asian case. *Special Issue on Forecasting in Management Science, 40*(6), 767–773. doi:10.1016/j.omega.2011.07.007.
20. Mahmoudi-Kohan, N., Moghaddam, M. P., & Sheikh-El-Eslami, M. (2010). An annual framework for clustering-based pricing for an electricity retailer. *Electric Power Systems Research, 80*(9), 1042–1048. doi:10.1016/j.epsr.2010.01.010.
21. Manera, M., & Marzullo, A. (2005). Modelling the load curve of aggregate electricity consumption using principal components. *Environmental Modelling & Software, 20*(11), 1389–1400. doi:10.1016/j.envsoft.2004.09.019.
22. McLachlan, G. J. (2004). Discriminant analysis and statistical pattern recognition. Hoboken: Wiley-Interscience.
23. Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K., & Kolehmainen, M. (2010). Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Applied Energy, 87*(11), 3538–3545. doi:10.1016/j.apenergy.2010.05.015.
24. Reichl, J., & Kollmann, A. (2010). Strategic homogenisation of energy efficiency measures: An approach to improve the efficiency and reduce the costs of the quantification of energy savings. *Energy Efficiency, 3*(3), 189–201.
25. Rousset, P. (1999). Applications des algorithmes d'auto-organisation à la classification et àă la prévision. PhD thesis, Université Paris I, Paris.
26. Soliman, S. Ah., & Al-Kandari, A. M. (2010). *Electrical load forecasting: Modeling and model construction.* New York: Elsevier
27. Taylor, J. W., De Menezes, L. M., & McSharry, P. E. (2006) A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting, 22*(1), 1–16.
28. Thang, K., Aggarwal, R., McGrail, A., & Esp, D. (2003). Analysis of power transformer dissolved gas data using the self-organizing map. *IEEE Transactions on Power Delivery, 18*(4), 1241–1248. doi:10.1109/TPWRD.2003.817733.
29. Tsekouras, G., Kotoulas, P., Tsirekis, C., Dialynas, E., & Hatziargyriou, N. (2008). A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers. *Electric Power Systems Research, 78*(9), 1494–1510. doi:10.1016/j.epsr.2008.01.010.
30. Vijayaraghavan, A., & Dornfeld, D. (2010). Automated energy monitoring of machine tools. *CIRP Annals Manufacturing Technology, 59*(1), 21–24. doi:10.1016/j.cirp.2010.03.042.
31. Vine, E. (2008). Breaking down the silos: The integration of energy efficiency, renewable energy, demand response and climate change. *Energy Efficiency, 1*(1), 49–63.
32. Vine, E. L., & Sathaye, J. A. (2000). The monitoring, evaluation, reporting, verification, and certification of energy-efficiency projects. *Mitigation and Adaptation Strategies for Global Change, 5*(2), 189–216.