Received: 24 June 2018 | Revised: 20 December 2018 | Accepted: 16 January 2019

DOI: 10.1002/pip.3117

WILEY PROGRESS IN PHOTOVOLTAICS

EU PVSEC PAPER

Modelling and forecasting PV production in the absence of behind-the-meter measurements

Tomas Landelius[1] | Sandra Andersson[1] | Roger Abrahamsson[2]

[1]Swedish Meteorological and Hydrological Institute (SMHI), Folkborgsvägen 17, 601 76 Norrköping, Sweden
[2]Tekniska Verken Linköping Nät AB, Brogatan 1, 582 78 Linköping, Sweden

Correspondence
Tomas Landelius, Swedish Meteorological and Hydrological Institute (SMHI), Folkborgsvägen 17, 60176 Norrköping, Sweden.
Email: tomas.landelius@smhi.se

Funding information
Energimyndigheten, Grant/Award Number: 43231-1

Abstract

This paper deals with hourly day-ahead prediction of the net electricity load at nine photovoltaic installations in a Swedish regional electricity network. The objective of the study was to develop, test, and evaluate a set of methods to predict the contribution of PV power to the grid without knowing the production and consumption "behind-the-meter." An indirect and a direct approach for prediction of the net load were evaluated. For the indirect approach, a model of the gross production was first estimated based on the open-source software PVLIB. The model was then used to predict the net load given a forecast of the gross consumption. Since we lacked a model of the latter, we used a "perfect forecast," in terms of measured gross consumption, to estimate the performance of this approach. In the direct approach, a model of the net load was estimated using either linear regression or an artificial neural network. Here, the model was used for prediction of the net load without any information about the gross consumption. Both approaches rely on information from a numerical weather prediction model together with net load measurements from the previous day. Forecasts using the indirect approach with perfect information about the gross consumption resulted in a normalized (with installed nominal power) RMSEn of 11%. The direct approach with the artificial neural network also resulted in an RMSEn of 11%, even though it did not have any information from behind the meter. Linear regression had an RMSEn of 12%.

KEYWORDS

artificial neural network, numerical weather prediction, PV power prediction, PV simulation

1 | INTRODUCTION

Meeting a growing worldwide demand for energy while addressing climate change is a challenge. One way of tackling it is to increase the generation of renewable energy, such as solar photovoltaic (PV) power. Renewable energy depends on weather conditions, and high-quality forecasts are required by transmission system operators, independent services operators, grid owners, and energy suppliers for managing the energy mix, maintaining grid balance, and trading on the energy market. As an example, reducing the day-ahead forecast error by just 0.1 percentage points is said to save the California independent services operator and California ratepayers more than \$2 million per year.[1]

Solar radiation fluctuates in time and space and is nontrivial to predict. For day-ahead forecasts, numerical weather prediction (NWP) models offer the best information.[2] In order to turn the solar radiation forecast into a forecast of PV production, information about PV panel orientation, installed nominal power, and geographical location

Nomenclature:
Acronyms: A, apartment block site; BTM, behind the meter; DNI, direct normal irradiance; ff10, wind speed at 10 m; GHI, global horizontal irradiance; H, household site; LL, forecast length; MEPS, MetCoOp ensemble prediction system; MetCoOp, Meteorological Co-operation on Operational NWP; NWP, numerical weather prediction; O, office building site; PV, photovoltaic; PVLIB PV, system simulation software library; RMSE, root mean squared error; RMSEn, RMSE normalized with installed capacity; t2m, temperature at 2 m; UTC, coordinated universal time;
Symbols: β, slope of PV module; γ, azimuth angle of PV module; γ_s, solar azimuth angle; θ_s, solar zenith angle; C, covariance matrix of input vectors to regression models; c_g^m, modelled electricity gross consumption; c_g^o, observed electricity gross consumption; D, training data set; d, day index; e_k, eigenvectors in principal component analysis; f^m, modelled feed-in (negative net load); f^o, observed feed-in (negative net load); m, mean of input vectors to regression models; n_D, number of days in the training data set; p_g^m, modelled PV gross production; p_g^o, observed PV gross production; r, Pearson's correlation coefficient; t, time (hours); u_k, eigenvalues in principal component analysis; w, parameter vector; x, input vector to regression models

is needed along with the NWP data. Many behind-the-meter (BTM) systems produce electricity primarily for on-site use before delivering excess energy to the grid, and often, only data about installed nominal power and the address is recorded.

Moreover, measurements of BTM PV generation are scarce. What is available is the measured net load, which does not equal actual electricity consumption/production since some portion of it may come from BTM solar PV generation. In Sweden, for example, it is only mandatory for the district system operator to make hourly measurements of the feed-in (negative net load). This causes a disconnection between the measured load and the actual electricity demand as predicted by models based on load measurements before the advent of noticeable amounts of solar PV. As an example, none of the California independent services operator load forecast models include the impact of BTM solar PV.[1] Re-estimation of these traditional load models without added explanatory variables will not solve the problem since they were not developed with BTM consumption in mind. This problem will be worsened given a growing portion of weather-dependent BTM battery storage and vehicle charging in combination with time-dependent electricity rates.

The aim of the study presented in the paper was to develop, test, and evaluate a set of methods for predicting day-ahead net load when BTM PV measurements are unavailable. It is a first step towards the goal to predict net load at different scales in a regional electricity network, eg, in terms of total production and production per entry point or per secondary substation. Previous work has tackled the problem with the general lack of BTM measurements of PV production either by trying to disaggregate the net load into gross production and gross consumption [3-5] or by upscaling model output from a few representative sites where BTM measurements are available (eg, previous studies[6-8]). Not much work has been done on the former approach as confirmed by others such as van der Meer et al[4] and Wang et al.[5] Finding representative sites as suggested by the latter approach is difficult in Sweden where there are no requirements on measuring the gross production, and the number of sites owned and supervised by the regional grid operators are few.

Here, we describe two ways to forecast the net load. The first one is an indirect method where net load is partitioned into gross production and gross consumption. The gross production was modelled with PVLIB Python,[9] developed at Sandia National Laboratories to simulate photovoltaic energy systems. It is an open-source software and can be downloaded from `http://pvpmc.sandia.gov/applications/pv_lib-toolbox`. The gross consumption can either be modelled with an existing load model, based on data prior to the PV installation, or with a new load model whose parameters are estimated together with the parameters of the PVLIB model. The latter method makes it possible to take into account possible changes in the consumption pattern following a PV installation.

The second approach is a direct approach where statistical models in terms of linear regression and an artificial neural network were used to estimate forecast models of the day-ahead net load. Neither information about gross production nor gross consumption was used in this approach. As a result, this model can be adapted to any changes in the consumption pattern following a PV installation. A possible drawback is that this approach does not provide information about the gross consumption, ie, the total energy demand at the site.

In Section 2, we present the data used for this study, and in Section 3, we go into detail concerning the estimation and use of the PV models for the two approaches mentioned above. The approaches were evaluated over an 8-month period, April to October, during which solar PV production is most relevant in Sweden. Results from this evaluation is presented in Section 4. The paper ends with Section 5 on discussions and conclusions.

2 | DATA

In this section, we describe the data used for the study. It consisted of two parts: measurements of energy production and consumption along with NWP forecasts for the corresponding dates. Both data sets consisted of hourly data covering the time period March to October 2016.

The data was divided into one training set, consisting of 80% of the data (140 days with 3360 hourly values per site), and one evaluation set with the remaining 20% (35 days with 840 hourly values per site). The division was made with a random sampling using a uniform distribution. This was done in order to end up with similar probability distributions for the time of year in the training and evaluation data sets.

2.1 | Measured production and consumption

The electricity measurements were provided by Tekniska verken, who are responsible for the electrical grid in the Municipality of Linköping, parts of the Municipality of Mjölby, and large parts of the Municipality of Katrineholm in Sweden. The data came from 220 PV installations connected to the grid; see Figure 1. Hourly measurements were made of the electricity transported to (feed-in) or from (net load) the grid. In this study, we made use of nine out of these installations that, in addition, have BTM measurements of hourly gross consumption and gross production. This allowed us to develop and evaluate methods for predicting the BTM PV power production as well as its contribution to the grid.

In this paper, we denote the gross consumption with c_g and the gross PV production with p_g. The feed-in is denoted with f, and here it can be positive or negative (indicating a positive net load). With this notation, we have, for any given hour t, that the feed-in is given by the difference between the gross production and the gross consumption:

$$f(t) = p_g(t) - c_g(t). \tag{1}$$

The measurements represent the mean values for the intervals $00 - 01, \ldots, 23 - 24$ UTC. The installations are located on the roofs of five households, four apartment complexes, and one office building. For each installation, information about the geographical location and the nominal installed capacity is known. No information is available regarding the slope and azimuth of the PV modules or if the panels are distributed on roofs in different directions. Such information could possibly be obtained from geographical information systems as the addresses are known. However, such information has not been considered here in order to keep the method more general.

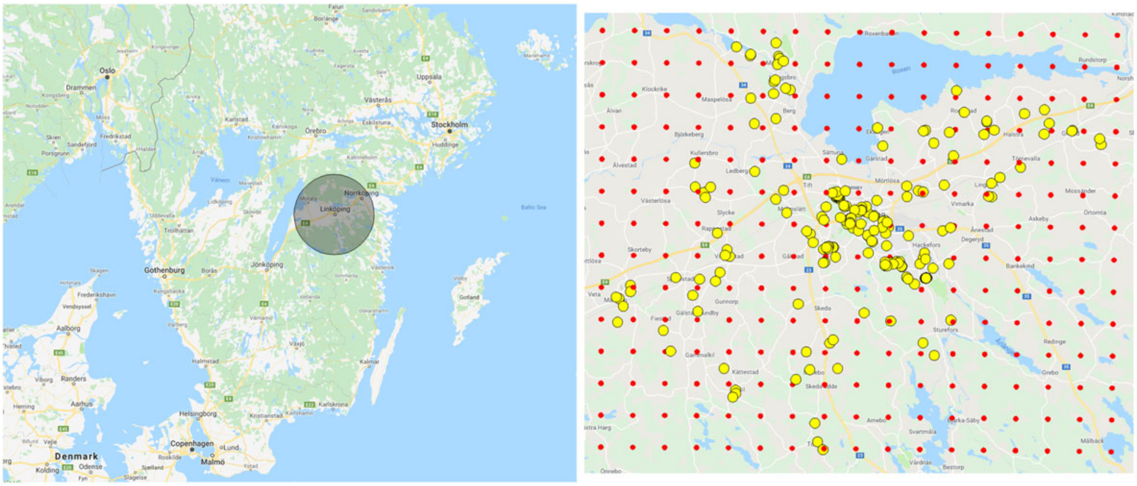

FIGURE 1 Left: approximate location of the Linköping network. Right: Numerical weather prediction (NWP) grid points (red) and photovoltaic (PV) installations (yellow) [Colour figure can be viewed at wileyonlinelibrary.com]

For this study, we only had data from sites with PV installations and only from the year 2016. Hence, we had no information about the gross consumption prior to the PV installations. Neither did we have access to historical data from the load model used by Tekniska Verken. The consumption pattern is different for weekdays and weekends. Here, we restricted ourselves to looking at weekdays. For the general case, another model is needed to cover weekends and holidays.

2.2 | NWP model data

NWP data was obtained from MetCoOp (Meteorological Co-operation on Operational NWP) where the meteorological services of Sweden, Norway, and Finland run the common ensemble prediction system MEPS (MetCoOp EPS).[10,11] The radiation model is based on the radiation scheme by Morcrette[12] and uses the Rapid Radiative Transfer Model.[13] The model domain contains 900×960 points with 2.5 km grid spacing (see example in Figure 1) and 65 levels covering a Nordic region. For this study, we only used the deterministic MEPS control forecasts started at 00 UTC with a length of 24 hours.

Hourly NWP data for the global horizontal (GHI) and direct normal (DNI) irradiances were obtained by taking differences between accumulated forecasts from $00 + LL + 1$ and $00 + LL$ where LL is a forecast length between 1 and 24 hours. Hence, forecasts for the time intervals $00 - 01$ UTC, …, $23 - 24$ UTC were obtained. The surface pressure, 2-m temperature (t2m), and 10-m wind speed (ff10) are available as instantaneous values. Here, we used the values assigned to the start of the interval. Hence, the measured power production and consumption refer to the same intervals as the GHI and DNI forecasts.

3 | LOAD FORECAST MODELS

The statistical load forecast models that are used in Sweden today are intended to describe the customers electricity consumption. With the introduction of more BTM solar PV, these models need to be extended to include PV electricity production and hence provide an estimate of the net electricity load.

Here, we studied two approaches towards this goal. The indirect approach complements an existing load model with a model for the BTM gross PV production. The net load is then given by subtracting the PV production from the original load model. The direct approach replaces the existing load model with a new one that also takes into account the weather variables that effect the consumption and PV production. As a reference, we also included a persistence model, ie, that tomorrow's hourly gross consumption pattern will equal today's.

3.1 | Indirect approach

The indirect approach is divided into two steps. First, a model of the gross production is estimated, and then this model is used together with a forecast of the gross consumption in order to end up with a forecast of the net load. A number of different ways to proceed with these two steps are described below.

3.1.1 | Estimation

For the indirect approach, PVLIB Python was used to model the BTM PV power production. This is an open-source community-supported tool for simulating the performance of PV energy systems. It was originally based on a toolbox developed at Sandia National Laboratories. For our purposes, we chose a module and inverter from the Sandia library in PVLIB that should match common installations in Sweden; `SunPower_SPR_220` and `ABB__MICRO_0_25`, respectively. Detailed data about these and other modules and inverters provided by Sandia can be found at `https://sam.nrel.gov/libraries`.

The gross production at each of the nine sites was modelled with a scaling factor times the PVLIB output from one module. This is an approximation since an installation may consist of modules in different orientations and may be affected by shading. We denote the PVLIB gross production model for a single site with $p_g^m(w, t)$, where the parameter vector w consists of the scaling factor together with the tilt and azimuth angles describing the orientation of the PV module. Besides these three estimation parameters, the PVLIB model also needs inputs in terms of global and direct normal irradiances, surface pressure, air temperature, wind speed, longitude, latitude, and time of day. All the weather-related information was obtained from archived NWP forecasts.

In order to estimate the parameter vector w, we need information about the gross production. Here, we propose the use of a cost function based on the daily load curve. This curve, with one mean value of the gross consumption for each hour of the day, should be easier to model than the full hourly time time series covering all 8 months. To our knowledge, this is a novel way to estimate a model for the gross production. Three estimation methods were studied:

1. The first method assumes that there is a daily load curve available from an existing load model or load measurements prior to the PV installation. This daily load curve is denoted $\bar{c}_g^m(t)$, where the bar indicates a mean value. The corresponding mean daily gross production curve from PVLIB is given by

$$\bar{p}_g^m(w,t) = \frac{1}{n_D}\sum_{d\in D} p_g^m(w,d,t), \tag{2}$$

where n_D is the number of days in the training data set D, and the hourly intervals refers to $t = 0-1, \ldots, 23-24$ UTC.

The parameter vector w is then estimated by minimizing the summed squared difference between the modelled and observed daily cycle of the mean hourly feed-in using a Nelder-Mead simplex algorithm (Python function `scipy.optimize.fmin`):

$$w^\star = \arg\min_w \sum_{t=0}^{24}(\bar{p}_g^m(w,t) - \bar{c}_g^m(t) - \bar{f}^o(t))^2. \tag{3}$$

Here, the subscript o is used to indicate the corresponding daily curve for the mean observed feed-in:

$$\bar{f}^o(t) = \frac{1}{n_D}\sum_{d\in D} f^o(d,t) \tag{4}$$

$$= \frac{1}{n_D}\sum_{d\in D}(p_g^o(d,t) - c_g^o(d,t)). \tag{5}$$

Note that in our case, we used observations of the gross consumption from the period March to October to calculate the daily load curve, ie, from the same time period as we have information about the net load. Hence, the results we present here based on this method will be better than what will be possible to achieve in the general case when information about gross consumption and feed-in will come from different time periods.

2. The second procedure is to look for the parameter vector that results in the smoothest daily load curve. This is a version of one of the methods from the paper by Sossan et al[3] where the smoothness is measured using a differentiated time series. Here, we differentiate the daily load curve:

$$\sum_{t=1}^{24}|\bar{c}_g^m(t) - \bar{c}_g^m(t-1)|. \tag{6}$$

We then want to minimize the sum in Equation 6 given our model of the daily load curve:

$$\bar{c}_g^m(t) = \bar{p}_g^m(w,t) - \bar{f}^o(t). \tag{7}$$

Hence, estimation of the parameter vector w means solving the following optimization problem:

$$w^\star = \arg\min_w \sum_{t=1}^{24}|\bar{p}_g^m(w,t) - \bar{f}^o(t) - (\bar{p}_g^m(w,t-1) - \bar{f}^o(t-1))|. \tag{8}$$

For this minimization, we need to employ an algorithm for constrained optimization since we want to restrict the parameter vector to regions where the resulting daily load curve is nonnegative. Another alternative is to add a penalty term to Equation 8 that grows large whenever the load curve becomes negative. In our case, we added a penalty term (a cost of 1000 for each negative parameter) and reapplied the Nelder-Mead simplex algorithm.

3. A third method is to describe the daily load curve with a parametrized profile generator or use principal component analysis to come up with a limited number of basis vectors that can describe the daily load curve. Samples of possible load curves for the principal component analysis could, for example, be obtained from load measurements prior to the PV installations or from load profile generators representing the sites under consideration.[14]

Here, we used measured daily load curves from the nine sites as samples and kept the first five principal components, $e_k(t)$, that explained about 90% of the sample variance. The daily load curve was then modelled as a linear combination of these five principal components:

$$\bar{c}_g^m(u,t) = \sum_{k=1}^{5} u_k e_k(t). \tag{9}$$

The vector u is then concatenated to the vector w (save the scaling parameter) describing the PVLIB model to solve the combined estimation problem as given by

$$(u,w)^\star = \arg\min_{u,w}\sum_{t=0}^{24}(\bar{p}_g^m(w,t) - \bar{c}_g^m(u,t) - \bar{f}^o(t))^2. \tag{10}$$

The reason for removing the scaling factor from the PVLIB parameter vector is to avoid redundancy since another scaling factor is introduced via the principal component coefficients in u. Again, constrained optimization or an additional penalty term is called for to ensure that the parameter vector u results in a model of the daily load curve that is nonnegative. In our case, we added the same penalty term and used the same minimization algorithm as described before.

3.1.2 | Forecasting

Once the optimal parameters of the gross production model have been estimated using one of the above methods, we can proceed and predict the feed-in given a forecast of the day-ahead gross consumption:

$$f^m(t) = p_g^m(w,t) - c_g^m(t). \tag{11}$$

We then face the problem of finding a forecast model for the gross consumption. A natural choice is to use an existing load model for this purpose. However, the drawback of such a procedure is that an existing model will not take into account possible changes in the consumption pattern after the PV installation.

Another way forward is to use the model for the gross production in order to find the corresponding gross consumption using Equation 11. Then, a new load model for the gross consumption can be estimated, including the effect of BTM PV generation. Such a procedure is however beyond the scope of this article.

For the present study, we instead used a perfect forecast of the gross consumption. This should provide us with a bound on the performance that is possible to achieve with the indirect approach. The perfect forecast is given by the measured gross consumption at the time when the forecast is valid.

3.2 | Direct approach

As an alternative to the indirect approach, we modelled the net load from the PV installation directly. This means replacing an existing, non-PV-aware, load model with a one that also takes into account the weather variables that effect the PV production. Here, the idea is to fit a parametric model to hourly data of the net load in order to get as many degrees of freedom as possible for the regression. We considered two regression alternatives: linear and nonlinear regression with an artificial neural network (ANN).

3.2.1 | Linear regression

The linear and the ANN model (to be described next) have different parameters but share the same input. The input vector, $x(t)$, for these two models consists of NWP forecast data for a given time tomorrow together with measured feed-in from the given time today:

$$x(d,t) = (\mathrm{GHI}(d,t),\ \mathrm{DNI}(d,t),\ \mathrm{t2m}(d,t),\ \mathrm{ff10}(d,t),\ f^{o}(d,t),\ \cos\theta_z(d,t), \\ \mathrm{GHI}(d+1,t), \mathrm{DNI}(d+1,t), \mathrm{t2m}(d+1,t), \mathrm{ff10}(d+1,t)). \quad (12)$$

The idea behind this set-up is that the model should be able to make a connection between today's weather and the feed-in. We also included the cosine of the solar zenith angle to provide some information about the time of the day and the season. It should also help calculating PV on tilted surfaces where the angle of incidence, θ, is related to the slope β and azimuth γ of the solar panel along with the solar zenith θ_z and azimuth γ_s angles[15]:

$$\cos\theta = \cos\theta_z \cos\beta + \sin\theta_z \sin\beta \cos(\gamma_s - \gamma). \quad (13)$$

For the linear model, we also added a constant to the input vector in order for the model to be able to add a bias. The linear forecast model is then given by

$$p_n^m(w, d+1, t) = w^T x(d,t). \quad (14)$$

We assume that the residual error is described by a normal distribution and employ a linear regression method to estimate the model parameters (Python function `numpy.linalg.lstsq`):

$$w^{\star} = \arg\min_{w} \sum_{d\in D} \sum_{t} (w^T x(d,t) - f^{o}(d+1,t))^2. \quad (15)$$

3.2.2 | Artificial neural network

Using machine learning to train nonlinear models has a long history within the area of energy forecasting.[16,17] Here, we used an off-the-shelf ANN from TensorFlow[18] to see if it could perform better than the linear model.

Using the standard feed-forward ANN estimator called `DNNRegressor`[19] from TensorFlow, we set up a network with a three-layer feed-forward topology with one input, one hidden, and one output layer. In our set-up, it had 11 inputs, 32, 64, or 96 nodes in the hidden layer, and one node in the output layer.

Determining the number of neurons in the hidden layer(s) is a trade-off between the networks ability to generalize from the training data (not too many neurons) and its representative power (not too few). Here, we were guided by the empirical relation for the number of hidden layer neurons proposed by Kalogirou[17]:

$$n_{hl} = 0.5(n_{in} + n_{out}) + \sqrt{n_{train}}. \quad (16)$$

Here, n_{in}, n_{out}, and n_{train} denote the number of input, output, and the size of the training data set, respectively. In our case, the number of inputs equaled 11, and we had one single output (the feed-in). The number of cases in the training data set was 3360. Hence, the suggested number of nodes in the hidden layer became 64. To check the robustness of this choice, we also tried with 32 and 96 neurons in the hidden layer. The network with 32 neurons actually performed slightly better than the others on the evaluation data set, so the results presented here are based on the outputs from that network.

In order to harmonize the amplitudes of the input variables, we removed the mean (m) and normalized the input vector by multiplying it with the inverse of the square root sample covariance matrix (C), calculated from the training data set. The TensorFlow minimization algorithm then found the solution to

$$w^{\star} = \arg\min_{w} \sum_{d\in D} \sum_{t} (w_o^T f(w_h, C^{-1/2}(x(d,t) - m)) - f^{o}(d+1,t))^2 \quad (17)$$

using an iterative procedure. The vectors w_h and w_o contains the weights for the hidden and output layers, respectively.

We ran the minimization for 10 000 iterations (saving the result at each 100th iteration) at which point the error for the evaluation data sets had started to increase for all sites. When this happens, the generalization capability of the ANN starts to deteriorate. For the optimal prediction network, we selected the parameters from the iteration for which the evaluation error had a minimum, ie, just before the ANN starts to perform worse on the independent data.

4 | RESULTS

In order to evaluate the performance of the different models, we computed some error measures. We compared the models by looking at the root mean squared error (RMSE) normalized by the nominal installed power (RMSEn), which is a common performance measure within the PV forecasting community.[6] We also calculated the square of the Pearson correlation coefficient (r^2) between the modelled and observed load. Only values when the sun was over the horizon were included in the calculations.

TABLE 1 Model performance in terms of mean r^2 values and the RMSEn for gross photovoltaic (PV) production day-ahead forecasts using PVLIB together with methods 1, 2, and 3 as well as using persistence at the different sites (mean value, H: household, A: apartment block, O: office). The RMSEn is the root mean squared error normalized with the installed capacity

		Mean	H1	H2	H3	H4	H5	A1	A2	A3	O1
PVLIB1	r^2	0.80	0.80	0.80	0.80	0.78	0.81	0.78	0.79	0.80	0.82
PVLIB2	r^2	0.76	0.79	0.79	0.71	0.78	0.80	0.77	0.80	0.80	0.57
PVLIB3	r^2	0.79	0.80	0.79	0.79	0.77	0.81	0.78	0.79	0.80	0.81
Persistence	r^2	0.45	0.46	0.45	0.48	0.42	0.44	0.45	0.43	0.43	0.46
PVLIB1	RMSEn (%)	11	11	12	10	11	11	11	11	12	9.2
PVLIB2	RMSEn (%)	13	11	13	13	11	11	11	12	12	20
PVLIB3	RMSEn (%)	11	11	14	11	11	11	11	12	12	10
Persistence	RMSEn (%)	21	18	22	17	29	18	18	29	21	17

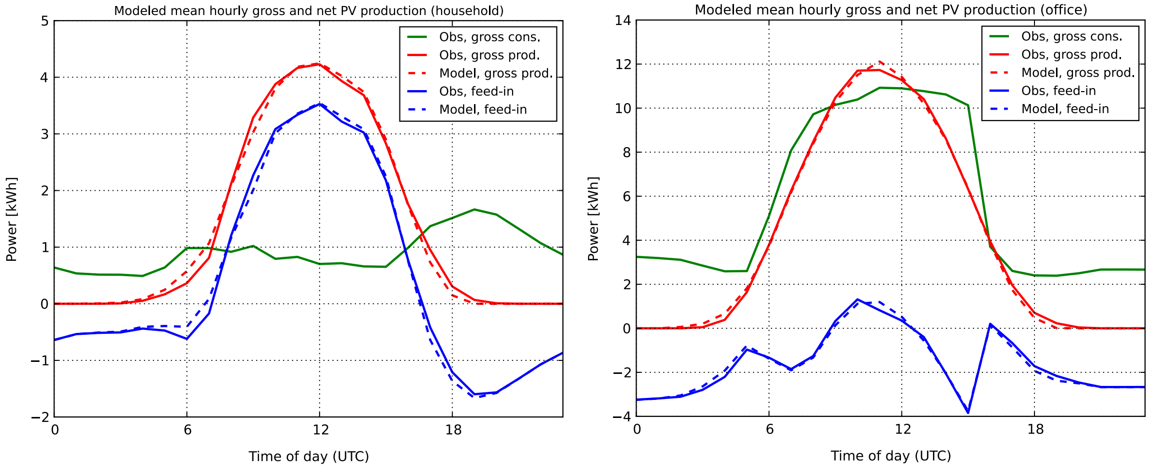

FIGURE 2 Estimation with PVLIB targeting the observed daily cycle of the feed-in (solid blue line) and the resulting model values (dashed blue line). The predicted (dashed red line) and observed (solid red line) gross production and observed consumption (green line) are also shown. Left: household. Right: office building [Colour figure can be viewed at wileyonlinelibrary.com]

First, we looked at the results for the estimation step of the gross production model with PVLIB for the indirect approach. Table 1 summarizes the error measures for the nine installations using the three different methods described in section 3.1.1. The average RMSEn of the resulting feed-in for method 1 that performed best were 11% compared with 21% for the persistence forecast. The error did not vary significantly, neither between the different installation types nor between the different methods. There was one exception though. Method 2 performed significantly worse for the office installation. The reason for this can be seen by comparing the daily load curves for the gross consumption (green lines) in the left and right panels of Figure 2. The assumption made using method 2 was that the load curve should be smooth and free from discontinuities. This was not the case for the load curve corresponding to the office installation. It shows steep steps during the morning and afternoon when the work at the office starts and ends. Method 3 employs a linear combination of principal components to model the daily load curve. It performed very similarly to method 1, which uses measured data to describe the daily load curve.

Figure 2 also shows two examples of using method 1 to fit the PVLIB model (blue dashed line) to the measured feed-in (blue solid line), which is the error criteria in Equation 3. Using the model to predict gross PV production (dashed red line) resulted in a good fit to observed values (red solid line). The daily pattern for gross consumption (green lines) of a household (left panel) shows the well-known structure with peaks during the morning and afternoon while the office installation (right panel) shows a consumption pattern related to the office hours.

The daily cycle of the mean RMSEn for the gross production forecasts are illustrated in Figure 3. Again, the PVLIB model for the gross production was estimated based on the gross production from the three different methods described in Section 3.1.1. The error corresponding to a persistence model is included to show the improvement over a naive approach. The error is shown as a function of forecast length. All forecasts were initialized at 00 UTC, and the PVLIB model outperformed persistence for all forecast lengths.

In the following, we used the model for the gross production estimated by method 1 as a reference when doing comparisons with the models based on the direct approach. Hence, the results from the indirect approach are based on perfect models of both the daily load curve and the hourly forecasts of the gross consumption. This should have provided us with an upper bound on the performance of the indirect approach.

Table 2 summarizes the error measures for the nine installations when using the indirect approach with PVLIB and the direct approach with linear regression and the ANN. A persistence model for the feed-in was again included as a reference. On average, the indirect

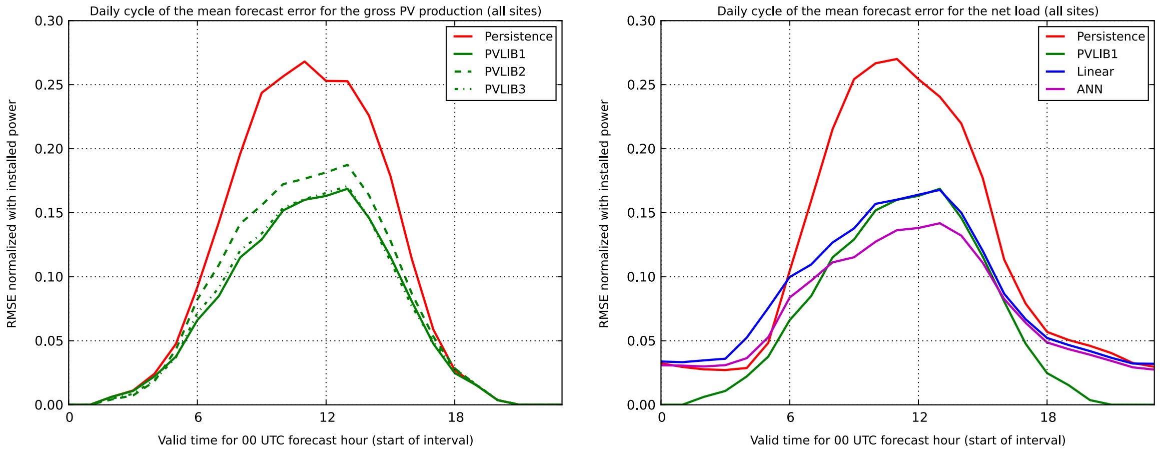

FIGURE 3 Left: Daily mean RMSEn cycle for gross photovoltaic (PV) predictions (left panel) and feed-in predictions (right panel). Estimation with PVLIB was performed using methods 1, 2, and 3. Net load predictions with PVLIB (method 1) were done using perfect forecasts of the gross consumption. ANN, artificial neural network; RMSEn, root mean squared error normalized with the installed capacity [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Model performance in terms of mean r^2 values and RMSEn for the feed-in using PVLIB (method 1), a linear model, an ANN, and persistence at the different sites

		Mean	H1	H2	H3	H4	H5	A1	A2	A3	O1
PVLIB1	r^2	0.79	0.83	0.80	0.80	0.80	0.82	0.71	0.80	0.80	0.71
Linear	r^2	0.71	0.69	0.78	0.74	0.78	0.78	0.60	0.76	0.77	0.49
ANN	r^2	0.76	0.74	0.81	0.77	0.82	0.81	0.65	0.82	0.81	0.63
Persistence	r^2	0.40	0.44	0.44	0.42	0.47	0.45	0.31	0.45	0.46	0.18
PVLIB1	RMSEn (%)	11	11	11	13	10	11	11	11	12	9.2
Linear	RMSEn (%)	12	13	13	12	11	12	12	12	12	12
ANN	RMSEn (%)	11	12	12	11	10	11	11	10	11	10
Persistence	RMSEn (%)	20	20	23	19	19	20	18	20	20	18

Abbreviations: A, apartment block; ANN, artificial neural network; H, household; O, office; RMSEn, root mean squared error normalized with installed capacity.

approach with PVLIB performed best with an r^2 value of 0.79 and an RMSEn of 11%. It was followed by the ANN and the linear regression model with r^2 values of 0.76 and 0.71, respectively, and similar RMSEn ratios of 11% and 12%. There was a slight difference in the r^2 values between different sites for the net load predictions. The correlation for the office site was notably smaller for all the models even though only weekdays are included in the data. This was to some extent also true for the installation on top of the apartment block A1. The reason why the net load at these two sites was harder to predict has not been investigated.

The daily cycles of the mean RMSEn when using PVLIB and the statistical models to forecast the feed-in are illustrated in the right panel of Figure 3. Again, the error is shown as a function of forecast length, and all forecasts are initialized at 00 UTC. Since the indirect approach based on PVLIB used a perfect forecast of the gross consumption, it resulted in a perfect fit during the night time. For the dark hours, the direct methods performed very similarly to persistence, indicating that this seems to be the information these methods rely on when no solar radiation is present. From late morning to early afternoon, the direct method with the ANN performed best. During most of the day, it can even perform better than the indirect approach based on a perfect forecast of the gross consumption. The error for the linear model was similar to that of the indirect method during the same time period. Both the indirect and the direct approach offered a substantial improvement over the persistence forecast between 06 and 18 UTC. For the linear regression and ANN models, the error was somewhat larger in the afternoon than in the morning. This can be explained by the NWP forecast deteriorating with the length of the forecast.

5 | DISCUSSION AND CONCLUSION

In this paper, we compared how a set of indirect and a direct approaches performed on the task of predicting the net load for the coming day at nine individual sites in a Swedish regional electricity network. The input to the presented forecast models consisted of information about measured feed-in from the previous day along with a NWP forecast for the next 24 hours. The fact that the study was done using data from Sweden should not severely limit the generality of the results. However, the results are probably of limited interest for countries or regions where BTM measurements are readily available. Moreover, in a real situation, the forecast has to be available well before the electricity market closes at about midday, and hence the forecast horizon needs to be stretched to at least 42 hours starting from 06 UTC. Such considerations will be the subject of further studies along with upscaling of the forecast to an area of a regional electricity network.

The indirect approach relies on a model of the gross PV production. Three methods for estimating such a model were described, all based on an auxiliary model of the daily load (gross consumption) curve. This is a novel approach as far as we know, and it allows us to come up with simpler models for the gross consumption than if all hourly values need to be described. The first method uses measurements of the gross consumption prior to the installation of the PV system. A drawback is that it will not capture any changes to the consumption pattern once the PV installation is in place. A second method is based on the assumption that the daily load curve should be as smooth as possible as suggested by Sossan et al.[3] This works well for the consumption pattern associated with households and apartments, but not for installations servicing offices with steep changes in the consumption connected to the office hours. The best performing method in the previous study[3] was based on separating production and consumption in the frequency domain. However, the introduction of batteries and time-dependant electricity tariffs will strive to make the consumption fit the production as closely as possible, rendering such a strategy less promising for the future. We instead propose a novel third method. Here, the principal component analysis is employed to describe the daily load curve as a linear combination of a limited number of basis function. The examples used for the principal component analysis can come from any installation types or be obtained from simulations with load profile generators. This provides a means for modelling load patterns that are not present in historical data from before the PV installation was made. Our evaluation showed that the third method performed o par with the first one even though that method was based on perfect information about the daily load curve.

References to other work on the indirect method are scarce as commented on in the introduction. Besides the publication by Sossan et al,[3] there is also work done by van der Meer et al[4] and Wang et al.[5] Van der Meer et al[4] are restricted to predictions for the next time step (with arbitrary resolution) based on production data at the current and previous time steps. Also, they don't use any exogenous input, like NWP forecasts, to explicitly take weather variability into account. Their methods are therefore not applicable to day-ahead power prediction where forecasts with hourly resolution can be required for the next 42 hours. In Wang et al,[5] the hourly net load is decomposed using an empirical model for the gross production and an ANN for the gross consumption. However, we question if this approach can be robust. In principle, an ANN can represent any function. Hence it should be possible to estimate ANN parameters so that the gross consumption match any gross production pattern associated with a given installed capacity and model orientation in the empirical model.

Any indirect approach needs to be complemented with a day-ahead forecast of the gross consumption in order to forecast the net load. In this study, we used a perfect forecast of the gross consumption. This means that both the model of the gross production and the gross consumption itself was based on perfect information. Hence, the performance of the indirect approach presented here should be seen as representing an upper bound. However, once a model of the gross production is in place, it can be used to obtain information about the gross consumption. For future work, this can in turn be used to estimate a model for the gross consumption.

For the direct approach, we used either linear regression or an off-the-shelf ANN from TensorFlow. The latter was included to see what kind of improvements a nonlinear model could offer. The direct approach estimates the model parameters based only on information about the hourly net load. It turned out that the performance of the linear regression model was similar to that of the indirect method based on perfect information about the gross consumption during most of the day. The direct approach based on the ANN even performed better than this indirect approach when the PV production peaks between 08 and 16 UTC. Averaged over all hours when the sun was above the horizon, the RMSEn for the best indirect approach using PVLIB and the direct approaches with linear regression and an ANN were 11%, 12%, and 11%, respectively. This is a substantial improvement over the RMSEn of 20% resulting from the baseline approach represented by a persistence forecast of the net load.

An ANN can theoretically represent any function, and hence it should be able to perform as well as or better than any indirect approach given that it is fed with sufficient amounts of data describing the problem. This line of reasoning seems to be in contrast with the results in Wang et al.[5] They compared an indirect and a direct approach, both based on ANNs with the same complexity, and got better results for the indirect method. A possible explanation could be that they in fact used a sequence of two ANNs for their indirect approach but only one for the direct.

Earlier studies have shown that the choice (eg, the ANN) between different nonlinear models is not critical.[20] However, better performance could perhaps be achieved if a recurrent structure is tried. Time correlations could then be exploited by allowing forecasted output up until time *t* to be used as input for the forecast ahead of time *t*. What to include in the input vector in general is another question for further investigations. Here, we picked information we thought was reasonable. No evaluation was made regarding how useful different input parameters were for the prediction. Future work should also look at using input from probabilistic NWP forecasts. This should be a way to describe and account for uncertainties in the solar radiation forecasts.

To conclude, we have presented a novel way to indirectly estimate a model for BTM gross PV production. We have also shown that forecasting the net load directly works as well, or during most of the day better, than forecasting the gross production and gross consumption separately. This is the case even if the forecast of the gross consumption is replaced with actual measurements.

ACKNOWLEDGEMENTS

The authors thank the reviewers for their feedback that helped improve the manuscript and also greatly acknowledge the Swedish Energy Agency that funded this work via grant 43231-1 as part of the national research programme "El från solen."

ORCID

Tomas Landelius https://orcid.org/0000-0002-3155-5696

REFERENCES

1. Monforte FA, Fordham C, Blanco J, Barsun S, Kankiewicz A, Norris B. Improving short–term load forecasts by incorporating solar PV generation. CEC-500-2017-031, San Diego, CA 92130: California Energy Commission; 2017.
2. Chaturvedi DK, Isha. Solar power forecasting: a review. *Int J Comput Appl*. 2016;145(6):28-50.
3. Sossan F, Nespoli L, Medici V, Paolone M. Unsupervised disaggregation of photovoltaic production from composite power flow measurements of heterogeneous prosumers. *IEEE Trans Ind Inf*. 2018;14(9):3904-3913.
4. van der Meer DW, Shepero M, Svensson A, Widén J, Munkhammar J. Probabilistic forecasting of electricity consumption, photovoltaic power generation and net demand of an individual building using gaussian processes. *Appl Energy*. 2018;213:195-207.
5. Wang Y, Zhang N, Chen Q, Kirschen DS, Li P, Xia Q. Data-driven probabilistic net load forecasting with high penetration of behind-the-meter PV. *IEEE Trans Power Syst*. 2018;33(3):3255-3264.
6. Elke L, Thomas S, Johannes H, Detlev H, Christian K. Regional PV power prediction for improved grid integration. *Prog Photovolt Res Appl*. 2011;19(7):757-771.
7. Sophie P, George G, George K. Solar and photovoltaic forecasting through post–processing of the Global Environmental Multiscale numerical weather prediction model. *Prog Photovolt Res Appl*. 2011;21(3):284-296.
8. Shaker H, Zareipour H, Wood D. Estimating power generation of invisible solar sites using publicly available data. *IEEE Trans Smart Grid*. 2016;7(5):2456-2465.
9. Holmgren WF, Andrews RW, Lorenzo AT, Stein JS. PVLIB python 2015. In: 2015 IEEE 42nd Photovoltaic Specialist Conference (PVSC); 2015; New Orleans, LA, USA:1-5.
10. Bengtsson L, Andrae U, Aspelien T, et al. The HARMONIE–AROME model configuration in the ALADIN–HIRLAM NWP system. *Mon Wea Rev*. 2017;145(5):1919-1935.
11. Müller M, Homleid M, Ivarsson K-I, et al. Arome-metcoop: A nordic convective-scale operational weather prediction model. *Weather Forecast*. 2017;32(2):609-627.
12. Morcrette J-J, Barker HW, Cole JNS, Iacono MJ, Pincus R. Impact of a new radiation package, McRad, in the ECMWF integrated forecasting system. *Mon Wea Rev*. 2008;136(12):4773-4798.
13. Mlawer EJ, Taubman SJ, Brown PD, Iacono MJ, Clough SA. Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated–k model for the longwave. *J Geophys Res-Atmos*. 1997;102(D14):16663-16682.
14. Grandjean A, Adnot J, Binet G. A review and an analysis of the residential electric load curve models. *Renew Sust Energ Rev*. 2012;16(9):6539-6565.
15. Duffie JA, Beckman WA. *Solar Engineering of Thermal Processes*. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2013.
16. Park J, Sandberg IW. Universal approximation using radial-basis-function networks. *Neural Comput*. 1991;3(2):246-257.
17. Kalogirou SA. Artificial neural networks in renewable energy systems applications: a review. *Renew Sust Energ Rev*. 2001;5(4):373-401.
18. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/, Software available from tensorflow.org; 2015.
19. DNNRegressor, a regressor for TensorFlow DNN models. https://www.tensorflow.org/api_docs/python/tf/estimator/DNNRegressor; Accessed: 2018-10-31.
20. Martin L, Zarzalejo LF, Polo J, Navarro A, Marchante R, Cony M. Prediction of global solar irradiance based on time series analysis: application to solar thermal power plants energy production planning. *Sol Energy*. 2010;84(10):1772-1781.

How to cite this article: Landelius T, Andersson S, Abrahamsson R. Modelling and forecasting PV production in the absence of behind-the-meter measurements. *Prog Photovolt Res Appl*. 2019;27:990–998. https://doi.org/10.1002/pip.3117