PV self-consumption prediction methods using supervised machine learning

Martos Tóth[1] and Nelson Sommerfeldt[1*]
[1]KTH Royal Institute of Technology, Stockholm, Sweden
**corresponding author: nelson.sommerfeldt@energy.kth.se*

Abstract

The increased prevalence of photovoltaic (PV) self-consumption policies across Europe and the world place an increased importance on accurate predictions for life-cycle costing during the planning phase. This study presents several machine learning and regression models for predicting self-consumption, trained on a variety of datasets from Sweden. The results show that advanced ML models have an improved performance over simpler regressions, where the highest performing model, Random Forest, has a mean average error of 1.5 percentage points and an R^2 of 0.977. Training models using widely available typical meteorological year (TMY) climate data is also shown to introduce small, acceptable errors when tested against spatially and temporally matched climate and load data. The ability to train the ML models with TMY climate data makes their adoption easier and builds on previous work by demonstrating the robustness of the methodology as a self-consumption prediction tool. The low error and high R^2 are a notable improvement over previous estimation models and the minimal input data requirements make them easy to adopt and apply in a wide array of applications.

Introduction

Prosumer photovoltaic (PV) policies are increasingly trending towards a self-consumption model (IEA, 2021), placing a high level of importance on predicting self-consumption for life cycle costing (Luthander et al., 2015). Self-consumed PV generation is the electricity used directly in the building, in contrast to the solar energy that is not needed and is sent/sold into the local grid. Since electricity meters take readings at discrete intervals, all self-consumption is a net value over some time period. Previous work has demonstrated the importance of interval on the final self-consumption value (Cao and Sirén, 2014), but due to their ready availability from utilities, hourly load profiles are still the most common (Sommerfeldt and Madani, 2017).

Early reports of measured self-consumption in Sweden show a high level of variance for a given ratio of PV generation to building load (Stridh, 2020), highlighting the difficulty and uncertainty associated with self-consumption prediction. In an ideal case, a specifically designed PV system's generation profile would be matched with an hourly load profile for making the life cycle cost calculations. However, in the planning stages or in urban energy models where specific load profiles are not available, a general purpose self-consumption model is needed.

Despite the increasing importance of self-consumption, there is relatively little empirical data or model development for making predictions. A study by McKenna et al. (2018), seemingly the first of its kind, used one-minute data from 218 homes in the United Kingdom to derive a linear regression model using annual PV generation (in kWh/yr) and annual electricity demand between 10:00 and 16:00 (in kWh/yr), finding a coefficient of determination (R^2) equal to 0.757. A subsequent study by Galli and Sommerfeldt (2021) built on this work by using a simulation approach to train several machine learning (ML) and regression models with hourly data from 108 Swedish villas over five years (2015-2019). The McKenna et al. approach was found to have an R^2 of 0.646, whereas the more advanced ML algorithms performed significantly better. The best results came from a Random Forest algorithm, producing an R^2 of 0.985 and a mean absolute error of 1.5 percentage points (of self-consumption).

The simulation approach of Galli and Sommerfeldt (G&S) includes some assumptions that limit the ability for comparisons to McKenna et al. First is that self-consumption is being calculated hourly instead of by the minute. While this is likely to lead to higher self-consumption values in general (Luthander et al., 2015), all values in the regression are summed to a total annual value, so the regression method is unaffected. Perhaps more relevant is G&S's use of the villa loads as generic profiles applied to typical meteorological year (TMY) data from several locations around Sweden. This disconnects the loads from the climate data, thereby potentially creating unrealistic self-consumption patterns, particular in villas relying on electric heating (e.g. heat pumps). This approach was used to mimic that which is most commonly applied by PV researchers and analysts, whereby one or several years of load data are considered representative of long-term patterns and is checked against the PV generation from a TMY climate file. However, this approach leaves open a question about model performance when the loads and PV generation are mismatched spatially and temporally.

Objective and Scope

This study builds on the work by Galli and Sommerfeldt by retraining the same ML regression models with spatially and temporally matched climate data. The objective is to quantify the error imposed by using TMY climate data combined with load profiles of a specific year to calculate PV self-consumption. The purpose of the models remains the same; with readily available annual data for any given location in southern Sweden, predict the PV self-consumption for use in a techno-economic optimization. This study will work towards validating the method and aims to provide insights into the variance of self-consumption year-on-year and its relevance for techno-economic PV analysis.

Methods

The hourly electricity loads are from the same 108 villas used by G&S from Karlstad, Sweden, taken from utility meters from January 2015 through December 2019. To quantify the impacts of spatial and temporal mismatch on self-consumption, three sets of regression models will be compared using various combinations of PV generation and self-consumption:

- The regression results from G&S, which were trained and tested using temporally and spatially mismatched climate data and loads,
- PV generation and self-consumption generated using TMY climate data from Karlstad, and
- PV generation and self-consumption generated using measured climate data from Karlstad between 2015 to 2019.

The first comparison is made between the regressions of each dataset trained and tested on its own predictions, to set a baseline variance between each approach. Then the models trained on the TMY climate data are tested against the Measured dataset from Karlstad, which quantifies the error imposed by using temporally mismatched data. Finally, an analysis of self-consumption values and their variance across time is presented to provide qualification to the results.

The ML regression models are accessed via the open-source SciKit-Learn libraries written in Python. Seven commonly used algorithms are tested, including; k nearest neighbours (k-NN), random forest, multi-layer perception (MLP), linear regression, polynomial regression, ridge regression, lasso regression, and the linear regression model proposed by McKenna et al. Model performance is measured using the mean absolute error (MAE), mean bias error (MBE), R^2 and R^2 adjusted.

The training/testing dataset consists of two inputs, annual PV generation (kWh/yr) and gross annual demand (kWh/yr), and one output, self-consumption as a unit less ratio between 0-1. G&S also tested tilt angle, azimuth, and latitude, however none of these inputs had a significant impact on the results. The k-fold cross validation method is applied (with k=10 folds) as well as feature scaling to give the inputs equal weighting. Complete details on the regression model parameters are described in (Galli and Sommerfeldt, 2021) and (Galli, 2021).

Building Loads

Building load data comes from 108 villas located in Karlstad, Sweden between 2015 and 2019. Figure 1 shows the annual electricity consumption of each villa and year, ranging from 1.5 MWh/yr up to 44 MWh/yr. The diversity of the dataset covers a continuous range of home sizes, occupants, and heating devices (none of which are known to the models), making it a valuable training set to represent the most populous areas of Sweden.

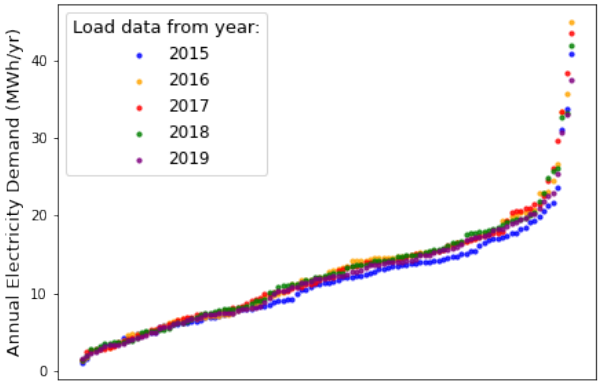

Figure 1: Distribution of annual villa loads

In addition to size and occupancy, seasonal load patterns vary by each load class due largely to the prevalence of electric heating and the northern location. Homes without electric heating have more consistent loads throughout the year, whereas those with heat pumps or direct elements see considerable increases in the winter months, as shown in the example from 2015 in Figure 2. These load patterns can have a significant impact on self-consumption due to the inversely proportional relationship between solar irradiance and space heating needs and is the primary motivation for using climate data that matches the loads.

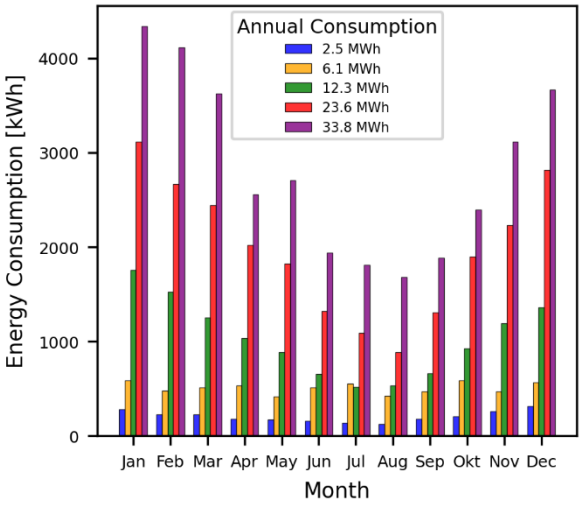

Figure 2: Monthly load pattern samples

PV Generation

In their study, G&S used TMY climate date from the PVGIS-SARAH database (Amillo et al., 2014). Here measured hourly data from the Swedish Meteorological and Hydrological Institute (SMHI) station at Karlstad Airport is used to create temporally and spatially matched PV generation to match the load data (SMHI, 2022). To test the impact of using spatially correct but temporally mismatched climate data, TMY climate data generated by Meteonorm 8.0 (Remund et al., 2020) is also tested using the years 1996 to 2015. Throughout the paper, these two datasets will be referred to as "Measured" and "TMY."

PV power generation is simulated using the PVWatts module in System Advisor Model 2021.12.02 accessed via PySAM 3.0.0 (NREL, 2022). The simulations are made with a 1 kW_p system using a typical performance ratio of 85% (Dhimish, 2020; van Sark et al., 2012), which is then scaled up or down in rated capacity to generate the self-consumption dataset. A full range of possible orientations are simulated with azimuth ranging from 0° to 300° in 60° steps, and tilts of 0° to 90° in 15° steps, resulting in 42 unique generation profiles. The yields (in kWh/kW_p) of each orientation from the TMY dataset are given in Figure 3 and monthly sums for each year from the Measured dataset in Figure 4. These results highlight the diversity of generation profiles used to create the self-consumption dataset.

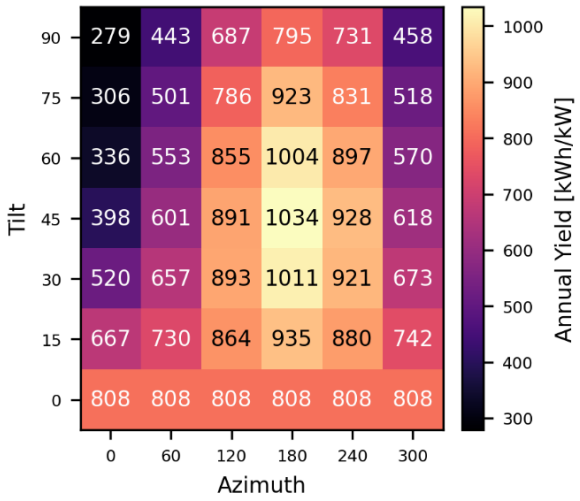

Figure 3: PV yield by orientation

In Sweden, and many other countries, prosumers are limited to PV system capacities that produce up to 100% of their annual load, effectively a net-zero building. Therefore each orientation's first year yield is used to determine a maximum capacity when paired with a building load and divided into 10 possible system sizes, i.e. scaling factors. The PV system parameters and their ranges are summarized in Table 1, which altogether result in 420 unique generation profiles applied to each building.

Table 1: PV system parameters

Parameter	Unit	Min	Max	Step
Capacity	% of Load	10	100	10
Azimuth	Degrees	0	300	60
Tilt Angle	Degrees	0	90	15

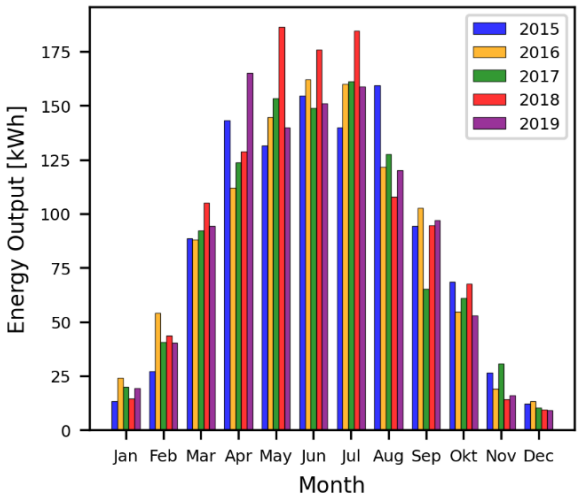

Figure 4: Monthly PV sums for 1 kW_p by year

Self-Consumption Data

Within each climate dataset, the 108 villas combined with 420 PV generation profiles over five years result in 226,800 unique self-consumption observations. To demonstrate the validity of the simulated self-consumption values, a comparison to published measured values are given in Figure 5. Self-consumption is shown as a function of solar fraction, which is the ratio of gross annual PV production to annual building load. Both the measured and simulated data come from 2018, with the blue points and fitted curve published by Stridh (2020). The average of all simulations are shown by the orange curve, with the maximum and minimum bounds shown with dashed black lines.

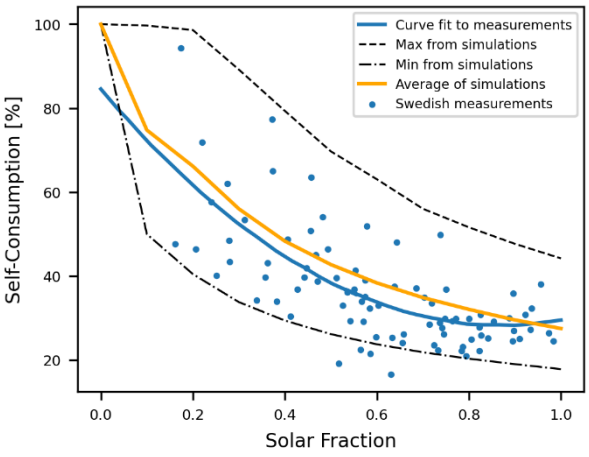

Figure 5: Simulated self-consumption from the 2018 climate data compared to measurements (Stridh, 2020)

The quantity of measured self-consumption data is relatively limited compared to what is possible with simulations, however the results in Figure 5 suggest the simulations provide a good representation of the real world. Aside from a few outliers, the max/min boundaries appear well placed and the average curves are within five percentage points across most of the data, oftentimes within just a few percentage points. One possible reason for the simulated self-consumption being higher than measured is the use of hourly data, as the published results are likely to be sub-hourly. However, the simulations presented here are notably closer to the measurements as compared to those by G&S, where the differences between the average curves were 5-10 percentage points, suggesting that the use of temporally and spatially matched PV generation will result in more accurate models. It is also worth noting that 2018 was a particularly sunny year in Sweden, which may also lead to lower self-consumption than a typical year.

Results

The results are presented in three sections; first are the regression results generated in this study as compared to G&S, followed by the test of TMY trained models on the Measured dataset, and finally the descriptive self-consumption statistics that help explain the regression model performance.

Updated Regressions

Like the regressions by G&S, the adjusted R^2 results are functionally identical (to within 0.001) to R^2, suggesting no overfitting of the models. Additionally the MBE results are nearly zero, with most models producing an MBE less than 1E-6 and the largest value being 0.001 from the MLP model. Therefore the adjusted R^2 and MBE are omitted from the tables, and the results comparison will focus on MAE and R^2, which are given in Table 2 and Table 3, respectively.

Table 2: MAE of Measured, TMY, and G&S regressions

Model	G&S	Meas	Diff	TMY	Diff
Rnd Forest	0.015	0.015	1.4%	0.015	-2.2%
k-NN	0.025	0.030	20.1%	0.030	18.7%
MLP	0.038	0.042	11.0%	0.042	11.0%
Polynomial	0.038	0.042	10.8%	0.041	9.1%
Ridge	0.038	0.042	10.9%	0.042	9.2%
Lasso	0.058	0.058	-0.3%	0.058	-0.4%
Linear	0.074	0.071	-3.4%	0.072	-2.9%
McKenna et al.	0.073	0.069	-5.4%	0.069	-4.9%

Both tables show that most of the models performed worse as compared to the original G&S results, however the range is highly varied. The largest differences come from the k-NN model, which has an MAE about 20% greater than G&S. The more basic regressions actually fared better with the new datasets, with up to 5% lower MAE in the case of the McKenna model. The rank order of model performance remains the same, with Random Forest still at the top and with only a marginal difference in MAE with either dataset as compared to G&S.

The R^2 results are similar in that most of the models perform slightly worse than the original G&S study. If using 0.9 as a general benchmark of performance, only the Random Forest and k-NN models are now above the mark. The underlying reason is unknown, but it is possible the k-NN model is more sensitive to input data volume and this revised dataset is smaller than G&S (226k vs. 1.08M). The best performing model, Random Forest, still retains a high R^2 and low MAE such that the quality of the model can be considered comparable to the original G&S models.

Table 3: R^2 of Measured, TMY, and G&S regressions

Model	G&S	Meas	Diff	TMY	Diff
Rnd Forest	0.985	0.977	-0.8%	0.980	-0.5%
k-NN	0.956	0.923	-3.4%	0.928	-2.9%
MLP	0.907	0.875	-3.6%	0.880	-3.0%
Polynomial	0.907	0.875	-3.5%	0.885	-2.5%
Ridge	0.897	0.857	-4.5%	0.868	-3.2%
Lasso	0.670	0.627	-6.5%	0.648	-3.3%
Linear	0.641	0.627	-2.1%	0.640	-0.2%
McKenna et al.	0.646	0.647	0.2%	0.658	1.9%

Performance Comparison

The next test uses models trained on TMY data with predictions applied to the Measured self-consumption dataset. The results, shown in Table 4, show the absolute values and the relative difference from the models trained on the Measured dataset. In every model the MAE increased by about 2% to as much as 40% in the case of the Random Forest. This large increase is in large part due to the low error values in general, which are still only 2.1 percentage points on average when applying the TMY model to the measured dataset. Another notable difference is the large increase in R^2 for the Lasso regression, which is an extreme outlier as compared to the marginal differences in the other models, but the cause is unknown.

Table 4: TMY trained models on Measured data

	MAE		R^2	
Model	Abs	Diff	Abs	Diff
Rnd Forest	0.021	40.0%	0.967	-1.0%
k-NN	0.032	6.7%	0.923	0.0%
MLP	0.044	4.8%	0.867	-0.9%
Polynomial	0.044	4.8%	0.869	-0.7%
Ridge	0.044	4.8%	0.868	1.3%
Lasso	0.059	1.7%	0.746	19.0%
Linear	0.074	4.2%	0.621	-1.0%
McKenna et al.	0.072	4.3%	0.641	-0.9%

Descriptive Statistics

The results of the previous two sections suggest that the ML method originally applied by G&S is robust. The differences in model performance with each training set is generally marginal, and cross-validating shows equally

small errors. This can in-part be explained by the dataset's variance across villas and years. Figure 6 provides a sampling of five villas, where the changes in electricity load year-on-year are visible. In most cases the changes are relatively small, but in Villas 1 and 3 there are differences of over 40% between the highest and lowest years. To capture a general overview of how loads vary over time, the relative standard deviation of all 108 villas is given in Figure 7, which normalizes the standard deviation to the average annual load and makes all villas comparable regardless of absolute annual load. This shows that for 75% of the villas, the change in load year-on-year will most likely be 10% or less. This first quartile shows a wide range of values, climbing as high as 0.48, which suggests a large change in activity or equipment (e.g. electric vehicle purchase or renovation).

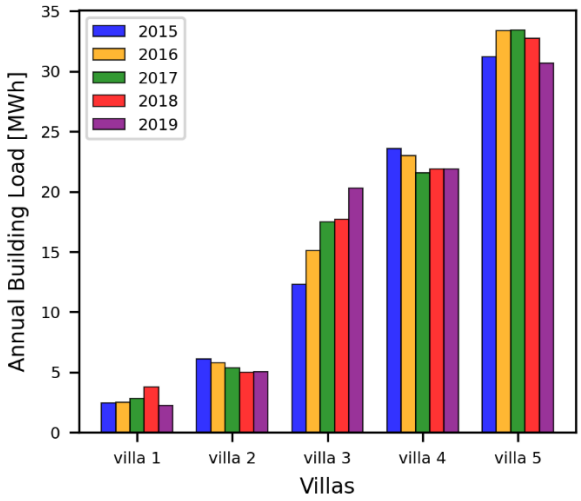

Figure 6: Sample of annual villa loads (MWh/yr)

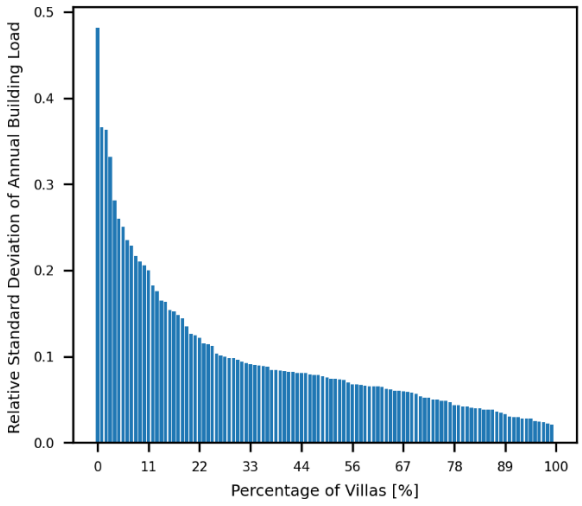

Figure 7: Relative standard deviation of annual load

A similar pattern for self-consumption is shown in Figure 8, where the year-on-year relative standard deviation for the Measured and TMY datasets are presented. Approximately 95% of all villas show a relative standard deviation of 10% or less, and about half are 5% or less. The TMY data, which has the same PV generation profile year to year, shows slightly lower variance in most buildings, but at most it is only a few percent. This helps to explain why the TMY trained models are still able to perform well against the Measured dataset. And while using load data from any given year is likely provide a suitable self-consumption prediction for use in life cycle costing, in some buildings or years this could lead to highly misleading results.

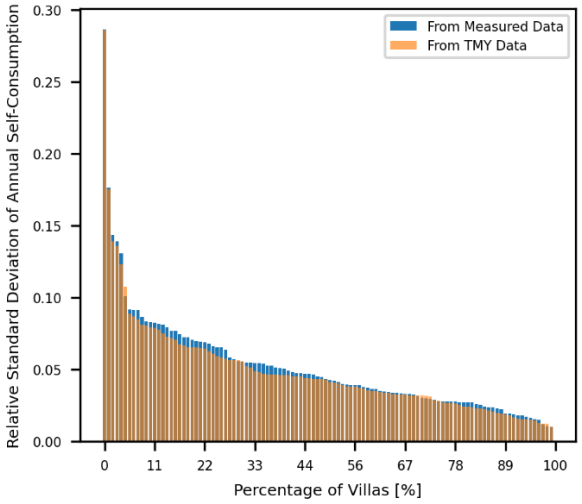

Figure 8: Relative standard deviation of self-consumption for all villas

Conclusions

This study builds on the work of Galli and Sommerfeldt (2021) and McKenna et al. (2018) by providing additional results for the training and testing of machine learning and traditional regression models to predict self-consumption.

The models trained using spatially matched (TMY) and spatially and temporally matched (Measured) climate data have marginally worse performance. In most cases, this difference is acceptable considering the low error values, particularly of best performing Random Forest model which has nearly identical performance with all datasets. The MAE of the Random Forest remained at 1.5 percentage points of self-consumption and the R^2 declined from 0.985 to 0.977 with the Measured dataset, still a notably high value.

To test the robustness of model training, the TMY trained models are tested against the self-consumption values of the Measured dataset. Here the performance of the Random Forest model is reduced, where MAE increases to 2.0 percentage points and R^2 falls to 0.967. Again, on their own these values are acceptable, particularly when considering the higher errors with traditional regression techniques. This result is encouraging for the spread of the model given that locally measured data is relatively difficult to acquire due to the lack of ground stations as compared to satellite derived TMY climate data (e.g. PVGIS or Meteonorm).

The detailed examinations on the variance of self-consumption over time help explain the low differences between TMY and Measured data and the high performance of the models. In most buildings, the change in self-consumption year-on-year is within a few percentage points, and given that the MBE values are nearly zero, this variance will trend towards zero over the lifetime of the PV system. The largest threat to accurate predictions is likely to come from significant changes in building load, for example from electrification of heating, an EV purchase, or the addition of batteries. A change in demand motivates a new self-consumption analysis, and future work should be directed towards training models that can represent these technologies.

From this study it can be concluded that the machine learning methods first presented by Galli and Sommerfeldt (2021) are robust and demonstrate an improvement over previous prediction methods, e.g. (McKenna et al., 2018). This technique adds value in that it only requires two, easy to acquire pieces of information; annual building load and annual PV generation. By comparison, many firms will use hourly load data, which in Europe is protected by GDPR and requires written permission from building owners to use. Alternatively, firms will also simply estimate self-consumption using past experience or published statistics. However, as Figure 5 shows, the variance for any given solar fraction is high, and the insights trained into the ML regression models reduces uncertainty by several percentage points.

This approach can also be valuable for more general and large-scale modelling tools. For example, the original G&S models are already deployed in publicly accessible solar maps, helping to automate the PV design process by optimizing system capacity with self-consumption.

While the methodology of using simulated PV data with measured loads is confirmed with this study, particularly given the improved representation of self-consumption using spatially and temporally matched climate data, it still does not constitute an empirical validation and should be performed in future work. It can also be interesting to train models with shorter time steps, particularly given the upcoming deployment of 15-minute metering throughout the country (Ei, 2017). It will also be interesting to test the method in other regions, where self-consumption patterns may differ and result in varying model performance.

Acknowledgements

The authors are grateful to the Swedish Energy Agency for funding this research via the Design for Everyday Energy Efficiency program, project number 48103-1. Thank you also to Fabian Galli for providing the original code this work is based on and to David Stoltz and Fredrik Balderud at Karlstad Energi for supplying the load data.

References

Amillo, A.G., Huld, T., Müller, R. (2014). A new database of global and direct solar radiation using the eastern meteosat satellite, models and validation. *Remote Sensing* 6(9), 8165–8189. https://doi.org/10.3390/rs6098165

Cao, S., Sirén, K. (2014). Impact of simulation time-resolution on the matching of PV production and household electric demand. *Applied Energy* 128, 192–208. https://doi.org/10.1016/j.apenergy.2014.04.075

Dhimish, M. (2020). Performance Ratio and Degradation Rate Analysis of 10-Year Field Exposed Residential Photovoltaic Installations in the UK and Ireland. *Clean Technologies* (2)2, 170–183. https://doi.org/10.3390/cleantechnol2020012

Ei (2017). Funktionskrav på elmätare (Functional requirements of electricity meters)

Galli, F. (2021) Predicting PV self-consumption in villas with machine learning. KTH Royal Institute of Technology.

Galli, F., Sommerfeldt, N. (2021). Predicting PV self-consumption in villas with machine learning. *38th European Photovoltaic Solar Energy Conference and Exhibition.* Lisbon (Portugal), pp. 993–997.

IEA (2021). Trends in Photovoltaic Applications (T1-41:2021).

Luthander, R., Widén, J., Nilsson, D., Palm, J. (2015). Photovoltaic self-consumption in buildings : A review. *Applied Energy* 142, 80–94. https://doi.org/10.1016/j.apenergy.2014.12.028

McKenna, E., Pless, J., Darby, S.J. (2018). Solar photovoltaic self-consumption in the UK residential sector: New estimates from a smart grid demonstration project. *Energy Policy* 118, 482–491. https://doi.org/10.1016/j.enpol.2018.04.006

NREL (2022). PySam Documentation. https://pypi.org/project/NREL-PySAM/ (accessed 1.31.22).

Remund, J., Müller, S., Schmutz, M., Graf, P. (2020). Meteonorm Version 8. *37th European Photovoltaic Solar Energy Conference and Exhibition*. Online, pp. 1466–1467.

SMHI (2022). Data. https://www.smhi.se/data (accessed 1.31.22).

Sommerfeldt, N., Madani, H. (2017). Revisiting the techno-economic analysis process for building-mounted, grid-connected solar photovoltaic systems: Part one - Review. *Renewable and Sustainable Energy Reviews* 74, 1379–1393. https://doi.org/10.1016/j.rser.2016.11.232

Stridh, B. (2020). Utvärdering av egenanvändning av solel i Sverige (Evaluation of self-consumption of PV electricity in Sweden). Swedish Energy Agency.

van Sark, W.G.J.H.., Reich, N.H., Mueller, B., Armbruster, A., Kiefer, K., Reise, C. (2012). Review of PV performance ratio development. *WREF 2012: World Renewable Energy Forum.* Denver (USA), pp. 4795–4800.