# Data-Driven Approach Utilising Random Forest Regression for PV Performance Monitoring

**Alexander David[1], Sabine Sint[1] and Thomas Bednar[1]**

[1] Research Unit of Building Physics, TU Wien, Vienna (Austria)

**Abstract**

The performance monitoring of PV plants is essential to ensure their correct operation, detect faults, and maximise their output. To be able to decide whether a plant is operating within normal parameters, a reference is needed, which indicates the PV performance that the plant should have under certain weather conditions. Weather data and historical monitoring data from times when the PV plant was operating correctly can be used to train machine learning models, which can provide the reference PV performance. In this research, a random forest regression machine learning algorithm is used to train models which predict the electrical power that is measured by 19 PV inverters and the PV main electricity meter. Several random forest models utilising different parts of the available weather data were trained. Their prediction performance was evaluated for five different time resolutions and three different PV module orientations. The results indicate that random forest regression is a suitable tool to predict the performance of PV plants.

*Keywords: photovoltaic, performance monitoring, performance prediction, machine learning, random forest regression*

## 1. Introduction

There are several methods and tools to calculate the performance of PV plants (Tozzi and Ho Jo, 2017; Umar et al., 2018). The science behind PV technology and its relation to weather data, geometrical parameters (e.g., orientation), and the surroundings (e.g., shading) is well-known (Mayer and Gróf, 2021). Even though a PV plant might have been planned in detail in a sophisticated PV simulation tool, usually, the PV plant owner does not have access to detailed planning information. It is common that the owner only receives a report indicating the expected PV performance for a year with average weather conditions. Such reports might be useful to evaluate whether the magnitude of the monthly PV production is in the correct range or not, but they are useless for monitoring the correct plant operation on an hour-to-hour or at least day-to-day basis. For monitoring a correct plant operation or its health status, there are various methods and algorithms (Pillai and Rajasekar, 2018), some even facilitating machine learning algorithms (Chen et al., 2018; Yao et al., 2021).

One relatively simple yet powerful machine learning algorithm is “random forest”. It can be used for classification and regression (Breiman, 2001). While it can excel at predicting data within the range of the data they were trained with, it fails when tasked to extrapolate beyond its trained scope (Breiman, 2001; Hastie et al., 2017). Extrapolation should generally not be required to predict the nominal PV electricity production at a given location as long as the training data includes weather that can typically be expected during a year. Thus, random forest models could be useful for fulfilling this prediction task.

In this paper, we evaluate the prediction performance a random forest regressor machine learning algorithm can achieve when predicting PV electricity generation at particular weather conditions. Instead of calculating the reference PV performance out of detailed plant configuration information, it is derived from historical data from when the plant was operating correctly. Thus, the PV plant owner does not need a detailed model of the system and simulation tool to simulate it but only sufficient monitoring data of the plant and weather data.

The data source used in this research is from the energy monitoring system of TU Wien’s (Plus-)Plus-Energy Office High-Rise Building. It is a highly energy-efficient building developed according to a net-zero energy concept. Within this concept, the primary energy source is a PV plant, which is integrated into the building’s


Selection and/or peer review under responsibility of Scientific Committee
doi:10.18086/eurosun.2022.16.03 Available at http://proceedings.ises.org

façade and mounted on its roof (Schöberl et al., 2014).

Fig. 1 shows the basic configuration of the building's PV plant. This graph summarises the information presented in (David et al., 2023). It illustrates the different PV sub-surfaces $A_i$ connected to one inverter $I_j$ and their PV module type. The stated surface sizes are the sum of the aperture sizes of the modules, and the stated power values are the sum of the modules' maximum power values at standard test conditions. The plant consists of four different parts: (i) southwest roof, (ii) southwest façade, (iii) southeast PV insulating glass, and (iv) southeast façade. Each illustrated combined surface is connected to one inverter – the only exception is inverter $I_1$, which is connected to the southwest-oriented surface $A_0$ and the southeast-oriented surface $A_1$. As indicated by Fig. 1, the plant's PV modules have one of these three orientations: (i) towards the southwest with a 15° inclination, (ii) towards the southwest with a 90° inclination, and (iii) towards the southeast with a 90° inclination.

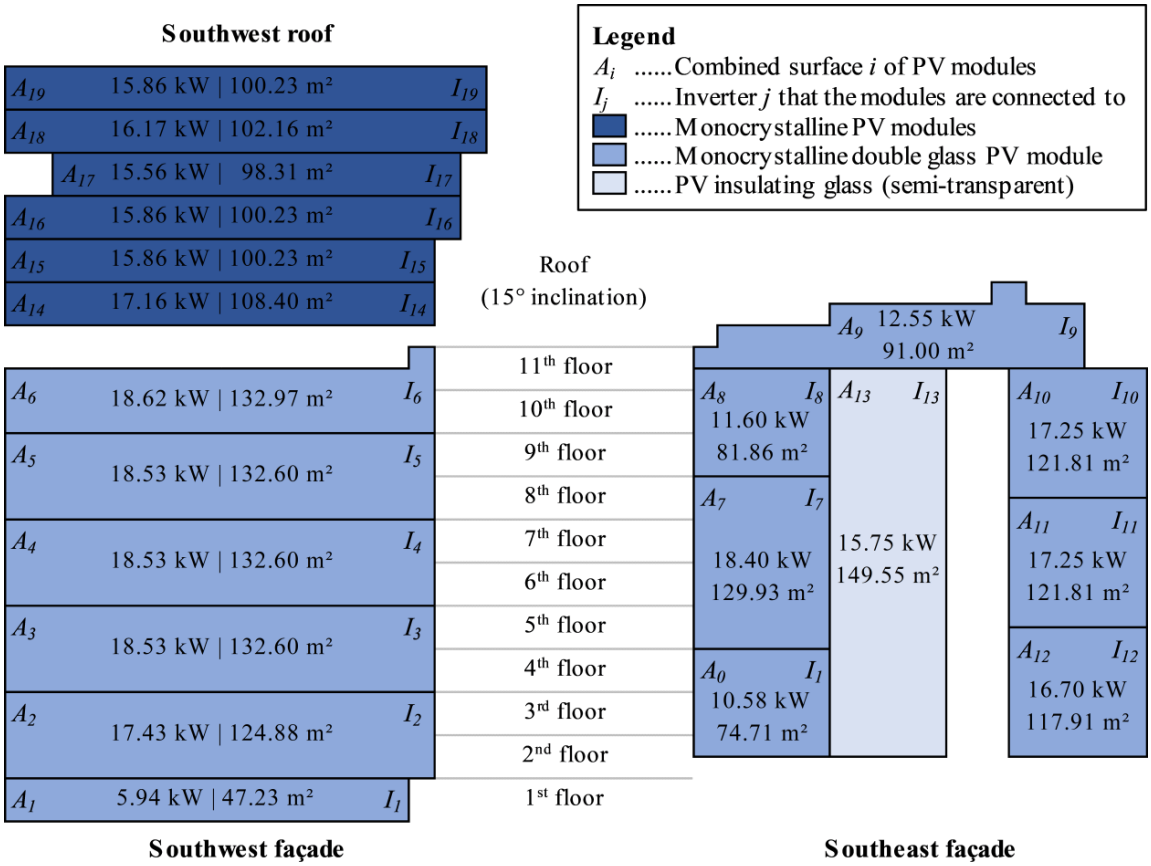


**Fig. 1: Overview of the (Plus-)Plus-Energy Office High-Rise Building's PV plant at TU Wien**

## 2. Method

This research aims to evaluate whether random forest regression is a suitable tool to predict the nominal performance of PV plants under certain weather conditions. For this purpose, historical weather data and PV monitoring data is needed from times when the PV plant operated as intended. This data must be split into a training and a test dataset – the training dataset for the training of the random forest models and the test dataset to evaluate their performance. This section shows how the monitoring data from the (Plus-)Plus-Energy Office High-Rise Building was processed and used to train and evaluate random forest models.

The (Plus-)Plus-Energy Office High-Rise Building is equipped with an extensive building energy monitoring system, which logs data from energy meters (electricity and thermal energy for heating and cooling), operational data (sensor data, setpoints and control signals), and weather data (external air temperature, global radiation, wind speed, …) in a 5-minute interval. One of the systems' electricity meters is the PV main meter, which meters the electricity production of the entire PV plant fed into the building's low-voltage main distribution. The electricity production of each of the 19 inverters is metered separately by integrated electricity meters. As every single meter from the 19 integrated ones and the PV main meter can be seen as the meter of a separate PV plant, they are treated

as such – each meter is used to train and evaluate a separate random forest model.

Since the integrated meters were initially not part of the building energy monitoring system, the data processed in this research is from when they were connected to the system, i.e., after January 2017. The monitoring data from the following two years was used for the training and the performance evaluation of the random forest models:

- Year 1: 04.02.2017 00:00 – 04.02.2018 00:00
- Year 2: 04.02.2018 00:00 – 04.02.2019 00:00

As the data from year 1 has several gaps with sizes ranging from a few days up to almost two weeks, and as year 2 is practically without gaps, year 2 was chosen as the training dataset and year 1 as the test dataset. All of the following processing steps described in this section were applied to both datasets.

Since inverters $I_3$ and $I_{14}$ had some failures during year 2, and as the goal of this method is to train random forests to predict the nominal PV performance that the inverters should have, their data was excluded from this research. The PV main meter is also affected by the failures of those two inverters, but as they are only parts of the entire plant, their impact was expected to be less significant.

For each meter, the time series of the load profile of the PV electricity generation is derived by calculating the first-order difference quotient of the time series of the respective meter's counter. The random forest models were trained to predict these load profiles as the target variables.

Since the models are intended to predict PV electricity generation under given weather conditions, the time series of some of the measured weather parameters are the predictor variables (short: predictors). Tab. 1 illustrates the weather parameters that were used in the training of the random forest models.

**Tab. 1: Parameters provided by the (Plus-)Plus-Energy Office High-Rise Building's weather station that are used as predictors in the random forest models**

| Measurement | Unit | Letter in the name of a random forest model if the model utilises this measurement for its predictions |
|---|---|---|
| (Global horizontal) irradiance | W/m² | i |
| (Air) temperature | °C | t |
| Relative humidity (of the air) | % | h |
| Illuminance on the north vertical surface | lx | n |
| Illuminance on the east vertical surface | lx | e |
| Illuminance on the south vertical surface | lx | s |
| Illuminance on the west vertical surface | lx | w |

Besides the weather parameters presented in Tab. 1, some further predictors were calculated out of time stamp information – Tab. 2 shows two time indicators that were used as predictors:

- The indicator "Sinus of the year" encodes the time of the year in the form of a sinus curve, which is fitted to the times when the solstices and equinoxes occur during a year. At the time of the summer solstice, this indicator is 1; at the time of the winter solstice, it is -1; and at the time of the equinoxes, it is 0.
- The indicator "Sinus of the day" encodes the time of the day in the form of a sinus curve, which is fitted to the times when noon and midnight occur during a day. At noon, this indicator is 1; at midnight, it is -1; and at 6 AM and 6 PM, it is 0.

**Tab. 2: Time indicators calculated out of time stamp information that are used as predictors in the random forest models**

| Time indicator | Unit | Letter in the name of a random forest model if the model utilises this indicator for its predictions |
|---|---|---|
| Sinus of the year | - | y |
| Sinus of the day | - | d |

One crucial part of developing machine learning models is the model selection. The goal is to train models that can accurately predict the target variables. To do that, they have to grasp the underlying connections between the predictors and the target variables based on the provided training datasets. When providing machine learning

algorithms with too many predictors, there is the danger of overfitting the models to the respective training dataset. To find out which predictors are useful for predicting the target variables and avoid overfitting, several models utilising different combinations of predictors must be trained. Their prediction performance is then evaluated by providing them with the test dataset. The decision of which model is best for the task at hand is then made based on the evaluation results (Hastie et al., 2017).

All the predictors in Tab. 1 and Tab. 2 have been assigned a single unique letter. In this paper, the designation of the developed random forest models is a combination of the letters of the predictors that each respective model utilises. For instance, if the model designation is "nesw", the model uses all four illuminance measurements from the weather data, and if the model designation is "ithydnesw", the model utilises all of the predictors presented in Tab. 1 and Tab. 2. The following eight model configurations were investigated in this research: ithyd, ithydnesw, itydnesw, itynesw, itnesw, inesw, tnesw, and nesw.

All of the models are set up as random forest regressors "sklearn.ensemble.RandomForestRegressor" from the Python package "scikit-learn 0.22.1" (Pedregosa et al. 2011) in the standard configuration with 100 estimator trees. As randomness is one of the key aspects of random forests, the models were repeatedly trained with the training dataset (100 repetitions). Their prediction performance at each repetition was evaluated with the test dataset. For each of the eight researched model configurations and each of the eighteen load profiles (PV main meter and all inverter meters except $I_3$ and $I_{14}$), a random forest regressor was set up by providing it with the respective load profile from the training dataset as target and the corresponding utilised predictors.

The prediction performances of all models were evaluated by providing the models with the predictors from the test dataset in order for them to predict the respective load profile. Each predicted load profile was then compared with the actual load profile in the test dataset. This was done by calculating the $R^2$ score and the deviation of the cumulated predicted load profile from the cumulated actual load (short: deviation).

With $n$ as the number of values in the time series of the load profile, $\hat{y}_i$ as the $i^{\text{th}}$ value of the predicted load profile, $y_i$ as the $i^{\text{th}}$ value of the actual load profile, and $\bar{y}$ as the mean of the actual load profile, the $R^2$ score is calculated as shown in eq. 1, and the deviation is calculated as shown in eq. 2:

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \qquad \text{(eq. 1)}$$

$$deviation(y,\hat{y}) = \frac{(\sum_{i=1}^{n}\hat{y}_i)-(\sum_{i=1}^{n}y_i)}{\sum_{i=1}^{n}y_i} \qquad \text{(eq. 2)}$$

While the $R^2$ score indicates of how well the course of the predicted load profile matches the course of the actual load profile, the deviation indicates the overall error that can be expected when predicting a whole year.

All of the steps described in this section were repeated for five different time resolutions: 5 min, 10 min, 15 min, 30 min, and 60 min. To accomplish this, the original energy monitoring data with a time resolution of 5 min was simply resampled to the other four time resolutions by utilising the method "pandas.DataFrame.resample" from the Python package "pandas 1.0.1" (McKinney, 2010) and calling its function "mean()".

As calculating PV electricity generation out of weather data is a problem with geometric transformations and strong linear dependencies, multiple linear regression should also be able to predict the eighteen load profiles accurately. Thus, a multiple linear regressor model utilizing the same predictors was also set up and examined for each of the eight investigated random forest model configurations. Due to the nature of multiple linear regression, their predicted PV electricity generation could be negative. To counteract this effect, negative values were always set to zero. The designation of the multiple linear regression models is identical to that of the random forest models but extended by the prefix "LR-".

## 3. Results

Even though in this research, the training and evaluation of regressor models were always conducted with the same training dataset and test dataset, the results from all repetitions, all time resolutions, and all model configurations are too numerous to be presented in detail. Thus, the results had to be condensed by grouping and averaging. To avoid that averaging of the achieved deviations cancels out the signs of the deviations, their average

is computed out of the absolute deviations.

Tab. 3 gives an overview of the average prediction performance ($R^2$ score and absolute deviation) that the eight different random forest model configurations achieved after being provided with data from the five different time resolutions. As the prediction results of the different load profiles of the inverters were quite similar if the orientation of their PV modules was the same, the results were grouped and averaged according to the PV module orientation. Since inverter $I_1$ is connected to PV modules with different orientations and as the PV main meter measures the sum of the load profiles of all inverters, they are displayed separately.

**Tab. 3: Average prediction performance that the random forest models achieved at 100 repetitions**

| Time resolution | Random forest model | $R^2$ score | | | | | Absolute deviation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean of | | | Inverter $I_1$ | PV main meter (sum of all inverters) | Mean of | | | Inverter $I_1$ | PV main meter (sum of all inverters) |
| | | South-west roof inverters | South-west façade inverters | South-east façade inverters | | | South-west roof inverters | South-west façade inverters | South-east façade inverters | | |
| 05 min | ithyd | 0.910 | 0.684 | 0.699 | 0.832 | 0.923 | 1.12% | 6.96% | 0.99% | 0.06% | 0.17% |
| | ithydnesw | 0.934 | 0.908 | 0.935 | 0.927 | 0.944 | 1.58% | 0.37% | 2.07% | 2.18% | 2.24% |
| | itydnesw | 0.933 | 0.908 | 0.934 | 0.926 | 0.944 | 2.38% | 0.78% | 2.40% | 2.85% | 2.59% |
| | itynesw | 0.932 | 0.907 | 0.932 | 0.923 | 0.943 | 2.46% | 0.79% | 2.65% | 3.19% | 2.85% |
| | itnesw | 0.926 | 0.905 | 0.932 | 0.918 | 0.941 | 2.65% | 0.61% | 2.44% | 2.96% | 2.57% |
| | inesw | 0.923 | 0.904 | 0.932 | 0.921 | 0.940 | 4.17% | 0.57% | 2.71% | 2.37% | 3.11% |
| | tnesw | 0.921 | 0.899 | 0.929 | 0.914 | 0.938 | 2.22% | 1.46% | 2.45% | 3.09% | 2.62% |
| | nesw | 0.916 | 0.897 | 0.928 | 0.914 | 0.936 | 3.53% | 1.35% | 2.67% | 2.35% | 3.05% |
| 10 min | ithyd | 0.943 | 0.754 | 0.788 | 0.882 | 0.957 | 0.33% | 5.53% | 0.75% | 0.83% | 0.76% |
| | ithydnesw | 0.957 | 0.935 | 0.956 | 0.950 | 0.968 | 1.23% | 0.54% | 1.59% | 1.69% | 1.83% |
| | itydnesw | 0.956 | 0.935 | 0.955 | 0.949 | 0.967 | 1.83% | 0.39% | 1.80% | 2.15% | 2.05% |
| | itynesw | 0.955 | 0.934 | 0.954 | 0.948 | 0.967 | 1.89% | 0.38% | 1.95% | 2.27% | 2.15% |
| | itnesw | 0.953 | 0.933 | 0.954 | 0.945 | 0.967 | 1.97% | 0.37% | 1.79% | 2.04% | 1.92% |
| | inesw | 0.949 | 0.932 | 0.953 | 0.946 | 0.965 | 4.01% | 0.34% | 2.24% | 1.86% | 2.71% |
| | tnesw | 0.950 | 0.931 | 0.953 | 0.943 | 0.966 | 1.16% | 0.35% | 1.72% | 1.89% | 1.71% |
| | nesw | 0.947 | 0.930 | 0.952 | 0.943 | 0.964 | 2.92% | 0.64% | 2.10% | 1.57% | 2.45% |
| 15 min | ithyd | 0.958 | 0.792 | 0.827 | 0.909 | 0.969 | 0.19% | 4.97% | 0.75% | 1.18% | 1.22% |
| | ithydnesw | 0.969 | 0.953 | 0.970 | 0.965 | 0.977 | 1.31% | 0.50% | 1.08% | 1.45% | 1.90% |
| | itydnesw | 0.969 | 0.952 | 0.969 | 0.964 | 0.977 | 1.75% | 0.31% | 1.21% | 1.84% | 1.97% |
| | itynesw | 0.969 | 0.952 | 0.969 | 0.963 | 0.976 | 1.75% | 0.29% | 1.27% | 1.88% | 1.97% |
| | itnesw | 0.967 | 0.951 | 0.968 | 0.961 | 0.976 | 1.78% | 0.34% | 1.25% | 1.89% | 1.81% |
| | inesw | 0.965 | 0.950 | 0.966 | 0.961 | 0.975 | 3.82% | 0.38% | 2.03% | 1.97% | 2.61% |
| | tnesw | 0.965 | 0.950 | 0.968 | 0.960 | 0.977 | 0.49% | 0.31% | 1.18% | 1.61% | 1.53% |
| | nesw | 0.964 | 0.949 | 0.966 | 0.960 | 0.975 | 2.42% | 0.60% | 1.91% | 1.58% | 2.26% |
| 30 min | ithyd | 0.978 | 0.854 | 0.891 | 0.945 | 0.982 | 0.56% | 4.71% | 0.65% | 1.55% | 1.55% |
| | ithydnesw | 0.986 | 0.976 | 0.984 | 0.981 | 0.987 | 1.20% | 0.68% | 1.11% | 1.21% | 1.87% |
| | itydnesw | 0.986 | 0.976 | 0.984 | 0.981 | 0.987 | 1.40% | 0.56% | 1.15% | 1.38% | 1.89% |
| | itynesw | 0.986 | 0.976 | 0.984 | 0.981 | 0.987 | 1.35% | 0.53% | 1.20% | 1.38% | 1.88% |
| | itnesw | 0.985 | 0.974 | 0.983 | 0.980 | 0.987 | 1.41% | 0.47% | 1.13% | 1.47% | 1.64% |
| | inesw | 0.981 | 0.973 | 0.981 | 0.978 | 0.986 | 3.89% | 0.29% | 2.09% | 1.78% | 2.40% |
| | tnesw | 0.984 | 0.974 | 0.983 | 0.978 | 0.988 | 0.06% | 0.48% | 1.07% | 1.26% | 1.29% |
| | nesw | 0.982 | 0.973 | 0.980 | 0.976 | 0.986 | 2.29% | 0.38% | 2.00% | 1.51% | 2.01% |
| 60 min | ithyd | 0.987 | 0.891 | 0.932 | 0.969 | 0.987 | 0.94% | 4.15% | 0.70% | 1.15% | 1.66% |
| | ithydnesw | 0.992 | 0.985 | 0.990 | 0.987 | 0.991 | 1.37% | 0.57% | 1.09% | 1.34% | 1.86% |
| | itydnesw | 0.992 | 0.985 | 0.990 | 0.988 | 0.992 | 1.48% | 0.49% | 1.03% | 1.42% | 1.82% |
| | itynesw | 0.992 | 0.985 | 0.990 | 0.988 | 0.992 | 1.47% | 0.48% | 1.05% | 1.42% | 1.81% |
| | itnesw | 0.992 | 0.984 | 0.989 | 0.987 | 0.992 | 1.41% | 0.29% | 1.01% | 1.52% | 1.69% |
| | inesw | 0.988 | 0.982 | 0.987 | 0.985 | 0.990 | 3.64% | 0.59% | 1.93% | 1.87% | 2.40% |
| | tnesw | 0.991 | 0.983 | 0.989 | 0.986 | 0.992 | 0.06% | 0.34% | 0.95% | 1.23% | 1.43% |
| | nesw | 0.988 | 0.982 | 0.987 | 0.985 | 0.990 | 2.05% | 0.64% | 1.86% | 1.45% | 2.13% |

As seen in Tab. 3, the average prediction performance is generally relatively high. Almost all models achieve on average $R^2$ scores around or above 0.9 and absolute deviations well below 3%. It can be observed that with lower time resolutions (larger time intervals), the prediction performance increases: At a time resolution of 60 min, the $R^2$ score almost always reaches values around 0.99, and the majority of absolute deviations stays below 1.5%. The load profiles of the southwest roof inverters and the PV main meter generally achieved the best prediction performance – even in the case of the model "ithyd", which has the poorest overall performance. The results indicate that the reason for it being the poorest is the absence of the illuminance values in the model's predictor variables.

The order of the random forest models in Tab. 3 is the order in which the model configurations were executed. While the model "ithyd" utilises the weather parameters usually found in weather data and the time indicators that can easily be calculated, the model "ithydnesw" is the same but extended by information about the illuminance on the vertical surfaces of the four cardinal directions. As these illuminance measurements are usually not part of the data provided by weather stations, the initial intention of this work was to develop random forest models without them. But as the prediction performance increased significantly when utilising the illuminance measurements, they were kept as predictors for all following models.

Since "ithydnesw" is the largest model configuration, utilising all predictors, it was assumed that it might be overfitted to the training dataset. To find out whether this is true and to detect another model configuration with a better prediction performance, the model was scaled down by leaving out some of the predictors. Although there are slight variations in the prediction performance, the following smaller models generally performed slightly worse than the full model "ithyndesw" – but their performance was still the same magnitude. This observation is also valid for the smallest model configuration "nesw", which exclusively uses the illuminance measurements as predictors. Since none of the models performs better than "ithydnesw", and its prediction performance is generally quite high, it appears that the model is not overfitted.

All models that utilise the illuminance measurements can generally be recommended for predicting load profiles – they achieve high performances over all of the investigated PV module orientations. As the high performances are achieved even though there are no direct illuminance measurements of the southeast and the southwest orientation, it is assumed that the load profiles of PV modules with orientations in the four cardinal directions north, east, south, and west will also be predicted accurately by those models.

The results indicate that the random forest model "ithyd" only achieves good prediction results if the load profile correlates with the global horizontal irradiance. In this research setting, this is only the case for the load profiles from the southwest roof inverters and the PV main meter.

Even though the load profile of the PV main meter is impacted by the failures of the inverters $I_3$ and $I_{14}$, all random forest regressors still achieved a remarkably high $R^2$ score and relatively low absolute deviation. This indicates that their impact on the load profile of the PV main meter is indeed not too significant. Nevertheless, it is assumed that the random forest models would have performed even better if there were no inverter failures.

Tab. 4 gives an overview of the prediction performance that multiple linear regression models achieve when provided with the same data as the random forest models. Even though all of the eight model configurations were evaluated in this research, only the three most distinctive cases are displayed in Tab. 4: a model utilizing all data except the illuminance (ithyd), the full model (ithydnesw), and a model using only the illuminance (nesw). All other models that are not explicitly displayed showed results that lay between the results of "ithydnesw" and "nesw". Generally, the results exhibited a pattern similar to the results of the random forests: "ithyd" is the model with the worst and "ithydnesw" is the model with the best prediction performance.

When comparing the results of the multiple linear regression models (Tab. 4) with the results of the corresponding random forest models (Tab. 3), it becomes obvious that the prediction performance of the random forests exceeded the prediction performance of the multiple linear regressions: The $R^2$ score was generally always higher and, in most cases, the absolute deviation was smaller. That seems to be because none of the provided predictors directly correlates with PV electricity generation. We expect the multiple linear regression models to achieve much better prediction performances when the irradiation or illuminance for the exact surface orientation of the PV modules is provided as a predictor.

**Tab. 4: Prediction performance that the multiple linear regression models achieved**

| Time resolution | Multiple linear regression model | $R^2$ score | | | | | Absolute deviation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean of | | | Inverter $I_1$ | PV main meter (sum of all inverters) | Mean of | | | Inverter $I_1$ | PV main meter (sum of all inverters) |
| | | South-west roof inverters | South-west façade inverters | South-east façade inverters | | | South-west roof inverters | South-west façade inverters | South-east façade inverters | | |
| 5 min | LR-ithyd | 0.884 | 0.546 | 0.510 | 0.750 | 0.865 | 1.73% | 19.22% | 5.15% | 4.73% | 3.17% |
| | LR-ithydnesw | 0.898 | 0.825 | 0.872 | 0.855 | 0.907 | 0.44% | 14.52% | 3.37% | 3.06% | 2.43% |
| | LR-nesw | 0.882 | 0.768 | 0.823 | 0.850 | 0.888 | 1.76% | 7.10% | 2.27% | 0.54% | 1.22% |
| 10 min | LR-ithyd | 0.921 | 0.566 | 0.532 | 0.786 | 0.899 | 0.72% | 18.73% | 5.06% | 4.21% | 2.46% |
| | LR-ithydnesw | 0.938 | 0.870 | 0.908 | 0.895 | 0.947 | 0.95% | 13.43% | 2.54% | 1.76% | 1.28% |
| | LR-nesw | 0.929 | 0.816 | 0.859 | 0.891 | 0.932 | 2.05% | 7.00% | 2.75% | 1.04% | 1.54% |
| 15 min | LR-ithyd | 0.938 | 0.576 | 0.543 | 0.802 | 0.910 | 0.32% | 18.53% | 5.09% | 4.19% | 2.26% |
| | LR-ithydnesw | 0.957 | 0.894 | 0.925 | 0.913 | 0.960 | 1.58% | 13.01% | 2.21% | 1.37% | 0.91% |
| | LR-nesw | 0.951 | 0.841 | 0.876 | 0.909 | 0.946 | 2.18% | 7.11% | 2.97% | 1.11% | 1.66% |
| 30 min | LR-ithyd | 0.956 | 0.590 | 0.556 | 0.821 | 0.922 | 0.29% | 18.36% | 5.10% | 4.18% | 2.11% |
| | LR-ithydnesw | 0.976 | 0.921 | 0.945 | 0.934 | 0.973 | 2.09% | 12.77% | 1.96% | 1.12% | 0.63% |
| | LR-nesw | 0.972 | 0.866 | 0.894 | 0.930 | 0.960 | 2.19% | 7.18% | 3.25% | 0.96% | 1.78% |
| 60 min | LR-ithyd | 0.962 | 0.600 | 0.563 | 0.833 | 0.927 | 0.40% | 18.22% | 5.12% | 4.23% | 2.02% |
| | LR-ithydnesw | 0.983 | 0.934 | 0.955 | 0.946 | 0.979 | 2.35% | 12.73% | 1.91% | 1.11% | 0.53% |
| | LR-nesw | 0.980 | 0.878 | 0.902 | 0.941 | 0.965 | 1.94% | 7.14% | 3.45% | 0.89% | 1.83% |

These results indicate that when only data from surfaces with a different orientation than the PV modules is available as predictors, the random forest regressors significantly outperform the multiple linear regression.

A detailed inspection of the prediction results of single repetitions of the random forest models shows that the average prediction performances presented in Tab. 3 correctly depict the performances to be expected in the different cases. Fig. 2 and Fig. 3 show the performance of three selected random forest models during one of the repetitions for a time resolution of 5 min. The prediction performance of the best multiple linear regression model (LR-ithydnesw) is also displayed in both figures. Since no averaging is involved, the deviation in Fig. 3 is presented as it is – i.e., it is not the absolute deviation as in Tab. 3 and Tab. 4.

When comparing the results presented in Fig. 2 and Fig. 3 with those in Tab. 3, it becomes obvious that the averaged results are very close to the results of one specific case. Thus, they are a reasonable way to represent the performance expected from the random forest regressors in the different cases.

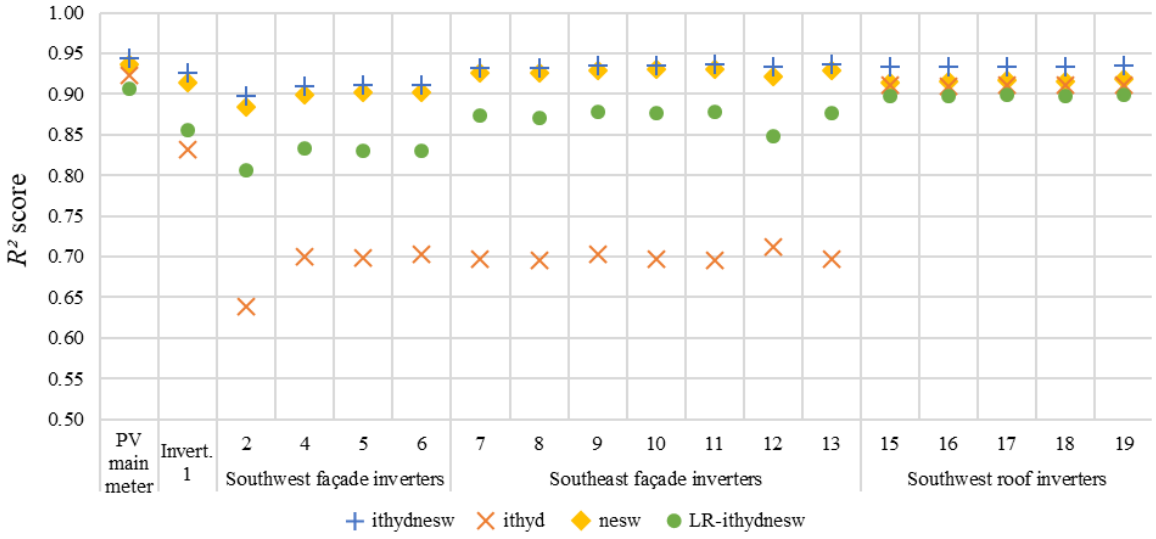


**Fig. 2: Prediction performance ($R^2$ score) that was achieved by three selected random forest models in the case of a 5 min time resolution during one of the repetitions and prediction performance achieved by the best multiple linear regression model**

Alike in Tab. 3, the performance of the models “ithydnesw” and “nesw” is remarkably high in Fig. 2: Regardless of the PV module orientation, the $R^2$ score lies between 0.9 and 0.95 for almost every load profile, and the absolute deviations are below 3% in most of the cases. Contrary to that, the model “ithyd” only performs well in the cases of the PV main meter and the southwest roof inverters. In the case of the façade inverters, its performance is significantly worse.

The best multiple linear regression model, “LR-ithydnesw”, never outperforms the random forest models “ithydnesw” and “nesw”. Except for the southwest roof inverters, this is also the case for the random forest model “ithyd”.

The depiction of the real deviation in Fig. 3 instead of the absolute deviation allows for additional observations: While the models “ithydnesw” and “nesw” generally underestimate the PV electricity production, the model “ithyd” generally overestimates it. As an underestimation could be explained due to the fact that the training dataset is from the year after the year that was used for the test dataset – the difference could (at least partially) be explained by the degradation of the PV modules. Further, in the case of the PV main meter, the difference most certainly is also caused by the failures of inverters $I_3$ and $I_{14}$ during 2018.

At the moment, there is no obvious, plausible explanation for the overestimation of the PV electricity production by “ithyd” – maybe it is simply an expression of its poor prediction performance in some cases.

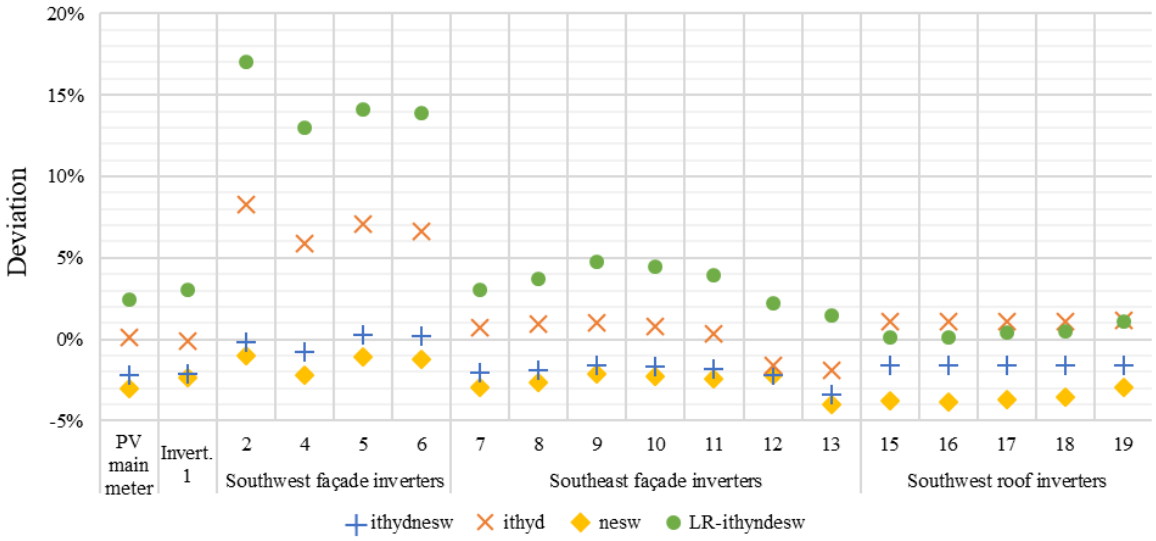


**Fig. 3: Prediction performance (deviation of cumulated predicted load profile from the cumulated actual load profile) that was achieved by three selected random forest models in the case of a 5 min time resolution during one of the repetitions and prediction performance achieved by the best multiple linear regression model**

All the results presented in Tab. 3, Tab. 4, Fig. 2, and Fig. 3 focussed on the performance indicators $R^2$ score and deviation of the cumulated predicted load profile from the cumulated actual load profile. To illustrate the underlying data and show which prediction accuracy can be expected during a day, Fig. 4 and Fig. 5 show comparisons of the time series of the predicted and actual load profiles. The presented time series are from the timespan 04.02.2017 00:00 - 09.02.2017 00:00, which is the very beginning of the test dataset. Again, the random forest models “ithydnesw”, “ithyd”, and “nesw” and the best multiple linear regression model “LR-ithydnesw” are shown in both depictions. The shown load profiles are only the profiles of inverter 7, one of the southeast façade inverters. This inverter was chosen on purpose, as in the case of the façade inverters, the model “ithyd” achieved a relatively poor prediction performance while the other models achieved a relatively high prediction performance. While Fig. 4 illustrates the case of a 5 min time resolution, Fig. 5 presents the case of a 60 min time resolution.

As seen in Fig. 4, the models “ithydnesw” and “nesw” could accurately predict the course of the actual load profile in the case of a 5 min time resolution. The predictions replicated the course of the PV electricity generation during a cloudy day (detail 1), its relatively smooth course during a sunny day (detail 2), and its more complex course on a sunny day with some clouds (detail 3).

These two random forest regressor models could accurately replicate the skewness of the daily course (caused by the southeast orientation of the inverter's PV modules) is remarkable – especially when considering the fact that none of the predictors directly correlates with the direct irradiation on a southeast-oriented surface. The random forest regressors appear to be able to infer the information needed to replicate such skewness from the predictors – assumedly by the irradiance measurements on the south vertical surface and the east vertical surface.

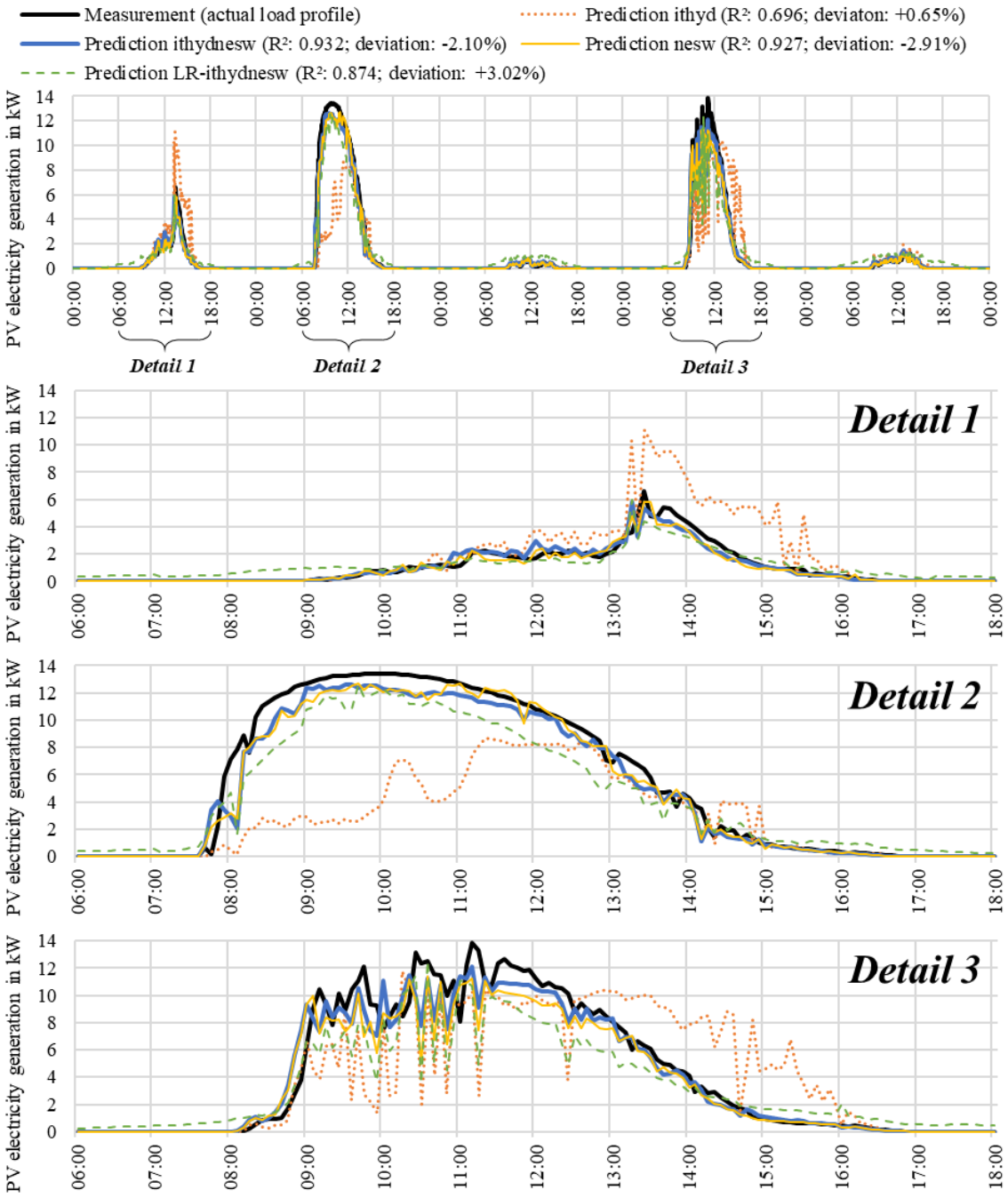


**Fig. 4: Comparison of the measured PV electricity generation of inverter 7 (southeast façade inverter) and the predictions from three selected random forest models and the best multiple linear regression model in the case of a 5 min time resolution (shown $R^2$ scores and deviations are the values of the predictions of a whole year)**

On the other hand, model "ithyd" predicted a load profile that is far off the real course of the actual load profile. The reason for that and the general poor prediction performance of "ithyd" seems to be that the information hidden in the "ithyd" training data is insufficient for the random forest model to correctly grasp the underlying connections between the predictors and the load profiles. It appears that the random forest regressors are simply

not able to reproduce the necessary geometric transformations out of weather data which only includes irradiation/irradiance values for one single surface orientation.

The load profile predicted by the multiple linear regression model "LR-ithydnesw" is also off the real course. It is generally too low during the day and too high during the night. Nonetheless, the prediction fits the actual load profile better than the prediction of "ithyd".

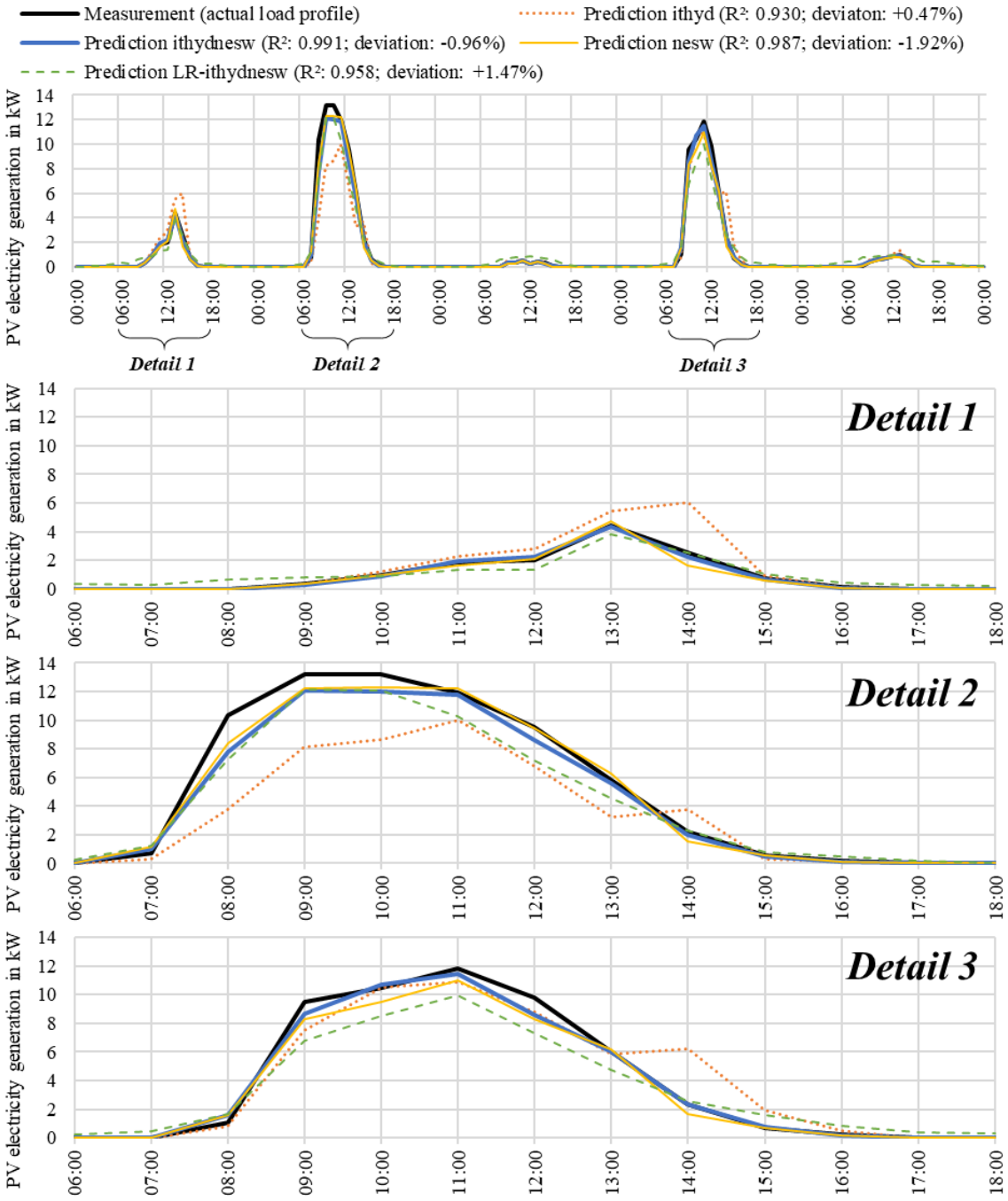


**Fig. 5: Comparison of the measured PV electricity generation of inverter 7 (southeast façade inverter) and the predictions from three selected random forest models in the case of a 60 min time resolution (shown $R^2$ scores and deviations are the values of the predictions of a whole year)**

For the case of a 60 min time resolution, see Fig. 5, the prediction performance of "ithyd" improved. Its prediction is much closer to the actual load profile, but it is still significantly worse than the predictions of the other three models. A direct comparison of Fig. 4 and Fig. 5 also explains the better $R^2$ score with decreasing time resolution: The load profile at a 60 min time resolution has a much smoother course than the one with a 5 min time resolution.

## 4. Conclusion

Given that the right weather data is available to be used as predictor variables, random forest regression is a suitable tool to accurately predict PV electricity generation. With increasing time resolution, the prediction performance generally gets slightly worse. The random forest regressors achieved remarkably high $R^2$ scores at a time resolution of 60 min. Thus, if the goal is to ensure the correct operation of the PV plant, predicting and comparing the PV electricity generation in a time resolution of 60 min can be recommended.

The best prediction performance can be expected from the full model ("ithydnesw"), which utilises all available weather information (horizontal global irradiance, air temperature, relative air humidity, and illuminance on the vertical surfaces of the four cardinal directions) and time indicators (yearly sinus function fitted to the solstices and equinoxes, daily sinus function fitted to noon and midnight). The model without the information about the illuminances ("ithyd") should only be used to predict PV electricity generation of plants where the load profile correlates with the global horizontal irradiance – this model cannot be recommended to predict load profiles of façade integrated PV modules.

The weather data used as predictors should include the illuminance measurements on the vertical surfaces of the four cardinal directions, north, east, south, and west, to achieve good prediction performances for all PV module orientations. In the case of the weather station used in this work, these parameters were not directly measured but instead calculated internally in the station itself. Since most weather stations do not provide illuminance on vertical surfaces, it is strongly recommended to calculate it using available weather data to use it as input for random forest regressors.

It is expected that if the illuminance is also calculated for all of the PV module orientations of a PV plant, multiple linear regression could also be generally recommended to predict PV load profiles. The downside is that at least the PV module orientations must be known, and the illuminance must be calculated specifically for each orientation.

The random forest regressors only need measurement data and the illuminance on the vertical surfaces of the four cardinal directions to achieve good prediction performances. The advantage is that the random forest regressors do not need specific knowledge about the PV module orientation predestines them for a large-scale application. They could be used in online monitoring systems provided by PV inverter manufacturers.

For such an application, the location of the PV plant must be known, and the weather data for this specific location must be obtained. The illuminance on the vertical surfaces of the four cardinal directions must be calculated from the weather data, and the sinus of the day and sinus of the year must be calculated from the time stamp information. After the monitoring data of the PV plant encompasses at least a whole year without uninterrupted operation, the full random forest regressor model ("ithydnesw") can be trained and then used to provide the reference for PV performance monitoring. To account for the degradation of the PV modules, the reference profile should be scaled down accordingly.

## 5. Acknowledgement

The authors acknowledge the BMVIT (Federal Ministry of Transport, Innovation and Technology), Kommunalkredit Public Consulting GmbH and MA20 (Vienna's municipal department for "Energy Planning") for funding this project. The authors also acknowledge all project partners, the TU Wien, the consortium of the Architects Hiesmayr-Gallister-Kratochwil, and the Bundesimmobiliengesellschaft as the building owner and operator.

## 6. References

Breiman, L., 2001. Random Forests. Machine Learning 45.1, 5-32.

Chen, Z., Han, F., Wu, L., Yu, J., Cheng, S., Lin, P., Chen, H., 2018. Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. Energy Conversion and Management 178, 250–264.

David, A., Sint, S., Bednar, T., 2023. Comparison of the Simulated and Measured Performance of the PV Plant

of Austria's Largest (Plus-)Plus-Energy Office Building. *EuroSun 2022: ISES and IEA SHC International Conference on Solar Energy for Buildings and Industry*, Kassel, 25-29 September.

Hastie, T., Tibshirani, R., Friedman, J., 2017. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition, Springer.

Mayer, M. J., Gróf, G., 2021. Extensive comparison of physical models for photovoltaic power forecasting. Applied Energy 283, 116239. https://doi.org/10.1016/j.apenergy.2020.116239

McKinney, W., 2010. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* https://doi.org/10.25080/majora-92bf1922-00a

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 2825–2830.

Pillai, D. S., Rajasekar, N., 2018. A comprehensive review on protection challenges and fault diagnosis in PV systems. Renewable and Sustainable Energy Reviews 91, 18-40. https://doi.org/10.1016/j.rser.2018.03.082

Schöberl, H., Hofer, R., Leeb, M., Bednar, T., 2014. Berichte aus Energie- und Umweltforschung, Schriftenreihe 47/2014: Österreichs größtes Plus-Energie-Bürogebäude am Standort Getreidemarkt der TU Wien. Vienna: Bundesministerium für Verkehr, Innovation und Technologie.

Tozzi Jr., P., Ho Jo, J., 2017. A comparative analysis of renewable energy simulation tools: Performance simulation model vs. system optimization. Renewable and Sustainable Energy Reviews 80, 390-398. https://doi.org/10.1016/j.rser.2017.05.153

Umar, N., Bora, B., Banerjee, C., Panwar1, B. S., 2018. Comparison of different PV power simulation softwares: case study on performance analysis of 1 MW grid-connected PV solar power plant. IJESI 7. 7, 11-24.

Yao, S., Kang, Q., Zhou, M., Abusorrah, A., Al-Turki, Y., 2021. Intelligent and Data-Driven Fault Detection of Photovoltaic Plants. Processes 9. 1711, 1-21. https://doi.org/10.3390/pr9101711