 Photovoltaic Energy Production Forecasting and Operational Analytics: A Real-World Study

Alexandros Alexakos†, Dimitrios Amaxilatis‡, Christos Zaroliagis† §
†Department of Computer Engineering & Informatics, University of Patras, Greece
‡SparkWorks Ltd, Galway, Ireland
§Computer Technology Institute & Press “Diophantus”, Patras, Greece
aalexakos@ceid.upatras.gr, d.amaxilatis@sparkworks.net, zaro@ceid.upatras.gr

Abstract—Photovoltaic energy production is one of the major parts for the shift of our societies away from fossil fuels. Their irregular nature however poses a challenge for electricity grid managers. PV energy production forecasting and monitoring is extremely important to guarantee that the grid will be able to supply consumers at all times and without any unexpected interruptions. Our goal is to provide build a system that can effectively monitor all aspects of the operation of a PV power station and provide owners, facility managers and power grid administrators with high quality forecasts and operational insights. To better understand and evaluate our work, we test our system in a real-world case study from Arta, Greece, and benchmark the effectiveness of both mathematical models and machine learning algorithms with regards to both precision, consistency and trustworthiness.

Index Terms—Monitoring, Internet of things, Embedded systems, Solar panels, Photovoltaic systems, Smart grids, Forecasting, Solar power output, Machine Learning, Random Forests, Feature selection, Prediction accuracy, Renewable energy system

I. INTRODUCTION

At a time when fossil fuels are dwindling, combined with global warming, a shift to green energy is imperative. The reckless use of mineral resources and combustion of fossil fuels results in rising global temperatures and worsening climatic phenomena. Renewable energy sources can help reduce the use of fossil fuels as an environmentally friendly alternative helping us reduce the effects of climate change. Especially in the evolving cities of the future, the use of renewable energy, with the installation of photovoltaic (PV) systems on roofs and urban hills can help not only benefit the environment but also increase their energy self-sufficiency. Limiting also the need to transfer high volumes of electricity from rural to urban areas can help cities overcome emergency situations when the electricity transfer lines may be severed or under maintenance. PV systems are also considered quite safe, both in terms of accidents at work (vs nuclear energy) and in terms of their recovery (vs mining) and do not contain any moving parts (compared to wind generators) that can be annoying people living nearby. Finally, they are economically advantageous as in addition to the cost of their installation, they do not require a large number of permanent staff for their operation, except for limited maintenance procedures.

Automating and simplifying not only the operation but the monitoring and the performance of PV systems can make their adoption even larger. To achieve this, with a minimal cost, we need to effectively monitor their operation and predict in advance any failures that can cause problems in their operation. Either placed in rural remote locations or in-city rooftops, remote monitoring technologies and data collection using cloud or edge computing and machine learning can make this work. Weather forecasts and sensors measuring the electrical properties of the PV panels can be used to extract predictions on the expected energy production and unexpected production fluctuations.

Forecasting energy production is also extremely important for grid managers and grid operators, as forecasting and nowcasting can help them anticipate the electricity generation from renewables for the coming minutes to the coming days and thus better prepare alternative sources to cover the gap with alternative fuel sources. Providing such estimations is also important for PV station owners. Accurate predictions can help them get better income, avoid fines in case of deviations or extended damages for surges that can be avoided even with limited attention and maintenance.

In this context, our goal in this work was to build a system for forecasting the electricity generated by PV power stations and identifying in-station incidents resulting in reduced production. We succeeded in meeting this goal and present here the system’s architecture and methods to detect all the aforementioned information from the power output, using machine learning and power production estimates to identify deviations, and classify them into specific malfunctions, in order to assist the PV operator identifying the source of malfunction and carry out the proper mitigation measures.

Our system operated for more than six months in a PV power station near Arta, Greece, in real world conditions, collecting data for the park’s power production and environmental conditions for the whole period. Our total dataset contains more than 2 million data points for more than 20 metrics regarding both electrical and environmental measures. The dataset is also augmented by the weather forecasts generated for each day in our dataset period (more than 180,000 data points for 10 parameters). Based on our results, we evaluate mathematical models and machine learning algorithms to generate next-day forecasts on the PV power production with a methodology that can be easily replicated for other installations. We also observe how our solution is capable of
generating insights on the operation of the PV installations based on the expected power production and actual production recorded on a daily basis.

The rest of the paper is organized as follows. In Section II we present the state of the art with regards to PV system monitoring and power production forecasting and analytics. Section III presents the testbed of our experimentation. The architecture of our system is depicted in Section IV with more insights on the data collection mechanisms and weather forecasting resources we use. In Section V we showcase the two forecasting techniques we implemented and evaluated as well as the results of our tests. Finally, in Section VI we present our conclusions and the next steps in our work.

II. STATE OF THE ART

Due to the challenge of climate and energy crisis, a lot of research has been devoted to both the monitoring and sensing infrastructures deployed in PV power plants and the forecasting and energy production techniques for estimating short or long time power generation.

Systems for intelligent remote monitoring of PV power conditioning units (PCU) and solar panels are showcased in [1] and [2]. The systems use embedded gateway devices that continuously poll the data either from the PCU or from sensors on the solar panels and record them in a cloud service using a GPRS connection. A similar system is presented in [3], where the authors use a custom device to measure a set of micro-inverters and collect the measured data in a cloud service. On the other hand, the authors in [4] present a system that monitors a PV power station at the panel level, using sensors to monitor temperature, humidity, irradiance, current and voltage. Such dense sensing infrastructures can provide owners with extremely useful information for the operation and performance of each individual panel, allowing for the detection of phenomena such as Potential Induced Degradation (PID) and Hotspots failures, but require a very high initial deployment and constant management cost to be sure that the collected data are reliable and not caused by a malfunction. The system presented in [5] uses a wide variety of sensors that identify parameters like temperature, humidity, irradiance, current and voltage outputs of solar panels but uses the collected data in conjunction with historical data to generate feedback for the levels of the energy generated and assess whether the facility needs an inspection, or maintenance due to a malfunction that has occurred.

On the domain of solar power forecasting, a comprehensive review of the theoretical forecasting methodologies for both solar radiation and PV power generation was provided in [6]. As they mention there are three major method categories used: **physical models**, that utilize solar and PV models to generate solar irradiance/power prediction, **artificial intelligence models**, like neural networks and **statistical models** that are based on historical data. A method that uses a machine learning model to predict energy and power was presented in [7], achieving a mean squared error (MSE) of 0.000000104, and a mean absolute error (MAE) of 0.00083. Three variations of a forecasting method that use a least-squares optimization of Numerical Weather Prediction are presented in [8]. These methods were tested in real world conditions in California for a period of 4 years (2011-2014) with a mean square error (RMSE) in the range of 10.3% to 14.0%. Finally, the case of how wind speeds can affect the performance of PV systems was presented in [9], with southerly winds providing from 20.4 to 42.9% increase in production when compared to northerly winds. Lately, [10] presented a solution for generating short-term energy production forecasts using deep learning and data from a PV installation in central Australia.

III. SOLAR FARM

The PV power station we study in this paper is located in Western Greece, in the broad area of city of Arta. It is composed of 4 independent electrical circuits that operate and produce energy independently. This separation is made for both managerial and operational / electrical reasons. More specifically it consists of:

- 3 subsystems of 248 rows of photovoltaic panels, with 24 panels in each row.
- 1 subsystem of 184 rows of photovoltaic panels, with 24 panels in each row.

In total, the park has 22272 panels with a power of 0.24 kWp, organized in 928 rows with a total power of 5345 kWp. The PV panels have been installed facing south and at an angle of 25° to the ground. In each of the first 3 subsystems, 4 inverters have been installed which are responsible for the conversion of direct current from the PV panels to AC, so that it can be fed to the power grid. In the fourth subsystem, 3 inverters have been installed for the same purpose. The inverters installed are Power-One: ULTRA-1100 and Power-One: ULTRA-1500\(^1\). A Feeder Protection Relay (ABB REF615) has been installed to connect the PV park to the electricity grid, which automatically controls the operation of the entire park.

The data for the production of electricity from the park is pumped through the Modbus interface of the Feeder Protection Relay. For the collection of meteorological data (panel temperature, ambient temperature, wind speed, solar radiation intensity) sensors have also been installed in the PV park which also work with the Modbus protocol. A Raspberry Pi device has also been installed in the park which continuously collects data locally from all Modbus interfaces (via network) and sends it to the central internet monitoring system (4G).

The software on the Raspberry Pi device can collect data while offline to minimize gaps in the collected data. The data collected offline is sent as soon as the system regains connectivity with the internet.

IV. SYSTEM ARCHITECTURE

Our system is split in two main parts. The first part contains all the on-site software and hardware components used to collect the sensor data from the PV power plant as well as

the cloud services that are used to store and maintain the data (on-site and edge layer). The second part is responsible for the communication with external APIs for collecting weather forecasts, forecasting of the power production as well as comparing on-site power production with the forecasted values for detecting unexpected deviations (cloud layer).

A. On-Site Data Collection

The data collection part of our system is designed upon the AWS IoT Greengrass\(^2\). The Greengrass deployment is comprised of multiple smaller applications, called components, that collect data independently from all the deployed sensors using the Modbus\(^3\) protocol over a wired network connection. Each component then sends its collected data to a local buffer module that stores the data temporarily in a SQLite database. Based on the connectivity status, the forwarder component picks the data from the SQLite database, and sends them to the cloud services of the system using the AWS IoT Core. A graphical representation of the Greengrass deployment is depicted in Figure 1. The system for the data collection has been in operation since late May 2021, and has operated without significant interruptions during the whole period, providing us with a large dataset of almost 6 months at the time of writing of this paper. This initial dataset consists of raw data, gathered with on site sensors, including solar and weather parameters (solar irradiance, outdoor temperature, wind direction and speed) and power generation parameters, regarding the operation of the PV panels (power generated, per phase voltage and current levels and alarms from the monitoring hardware). The whole dataset has a total size of 137 MB, with a total of 2,456,418 measurements. To cleanup the collected data, a simple rule was used to remove data recorded during the night when sun light was absent and the PV panels were not generating any power. When we aggregate the available data in an hourly basis a more concise dataset is extracted with data for 4,260 hours (almost 6 months of data).

![Fig. 1. The software components in the on-site gateway device of the data collection AWS Greengrass deployment](image)

B. Cloud Data Storage

The data collected from all the on-site sensing infrastructure is stored in Amazon Timestream\(^4\). Timestream is a time series database that gives us the flexibility to store large data volumes with high performance and low cost. It also provides us with an interface for easily executing aggregation, projection and gap filling operations on the database itself solving a lot of underlying problems when accessing previously collected data. Data can be extracted from the database using a provided API and converted either in json, csv or other formats as needed by each application developed.

A second data source for our system is an online weather forecasting service, more specifically tommorow.io\(^5\). This service was found to be the most accurate and reliable for the location of the PV power plant we studied and provided an easy to use API for accessing daily forecasts. To execute this process automatically, a scheduled job runs periodically (every 24 hours), retrieves the full weather and solar radiation forecast for the next 24 hours and stores them temporarily in an AWS S3 bucket, for each of the monitored locations. We are then able to download all forecasts for the expected period of our study and merge them into our dataset. These data are then used to predict the power output of the PV power plant that we use in our analysis. The weather forecasts, are also compared to the ones from the on-site sensors to check the forecast quality and better understand the expected power forecast deviations computed later on.

C. Data Comparison

As a method to verify the operation of our on-site sensors and assess the quality of the forecast data provided by the tommorow.io service, we compare their hourly averages for three main data types: solar irradiance, temperature and wind speed. Figure 2 shows the comparison of the average hourly data for the whole dataset collected thus far. We observe that the irradiance data are quite close, with a difference in hour 0 (caused by on-site lights affecting the sensor data) and the early hours of the day, where we observe that the irradiance recored by on-site sensors is lower than the forecast. On the other hand the irradiance forecast has a lower maximum value at the peak of the day. Small deviations are also observed with regards to temperature and wind speed. These are probably related to geography of the area, as there are two hills surrounding the installation from the north and the south, creating a slightly different micro-clima that can affect wind speeds and air temperature. Also the on-site sensors offer us a better understanding of the current conditions while weather forecasts target the wider area of Arta, and are probably generated using an approximation between the installed weather stations and forecast models of the weather services used.

V. POWER PRODUCTION FORECASTING

In order to provide the appropriate power production forecasts, our system uses the data from the two data sources

\(^2\)https://aws.amazon.com/greengrass/

\(^3\)https://en.wikipedia.org/wiki/Modbus

\(^4\)https://aws.amazon.com/timestream/

\(^5\)https://www.tomorrow.io/
Fig. 2. Average Irradiance, Temperature and Wind Speed recorded with on-site sensors and forecasted by tommorow.io for every hour of day for the whole dataset period. Each forecast, covers a period of 24 hours for which the system can generate an estimate for the power production. The power production is computed using 2 different methodologies. The first method uses physical models to compute the power production over the next 24 hours using the description of the power station and the irradiance levels forecast for the next 24 hours. The second uses the real data collected from the station over a large period to train a machine learning model to forecast the power production using no additional information for the actual park’s setup.

A. Physical Model Based Forecasting

For the physical model forecasting, we use the pvlib library [11]. Pvlid is a community supported tool that provides a set of functions and classes for simulating the performance of photovoltaic energy systems, originally ported from the PVLIB MATLAB toolbox developed at Sandia National Laboratories. In order to compute the forecasted power production the PVlib needs a set of information for the monitored installation including: the type of inverters installed, the number of solar panels used, their organization inside the power station (e.g., number of panels per inverter), their orientation (azimuth) and the tilt from the ground. This data is used to generate a ModelChain of the power station. The weather forecast is then fed to this ModelChain in order to generate the expected power production for the forecast period.

Based on the constructed model and the forecasted irradiance and weather data, we compute the expected generated power. By analyzing the results, we get a root mean square error (RMSE) value of 710. A high error value is to be expected based on the comparison we saw in Figure 2. One of the main parameters used for calculating the power production forecast is the irradiance levels. The actual irradiance levels are in general lower than the forecast during early morning due to the hills surrounding the installation and as a result the actual power production during these hours is lower than the forecasted one. Additionally during peak hours we see that the forecast is slightly lower than the actual irradiance recorded on-site. This can be again due to the inaccuracy of the forecast service for our specific location or to the increased diffuse radiation generated by reflections on the solar panels and the hills, leading to radiation intensity reaching the panels.

The average power production recorded with the on-site sensors and the average forecasted power production for every hour of the day is presented in Figure 3. In the same figure we also present the mean root square error for every hour of the day. As expected the error is higher for the first hours of the day.
day while it is reduced after 6 in the morning. This error is in fact expected at some extent as we saw that the forecasted irradiance is lower than the actual values recorded on-site. To tackle this forecast inefficiency, and as we are unable to get more exact forecasts for every possible location, we can apply a transformation on the forecast data, using historical data to generate an adjustment factor that can bring the forecast data closer to the actual on-site measurements.

B. Machine Learning Based Forecasting

For the machine learning based forecast we use the Random Forest algorithm [12]. The data we feed to the Random Forest algorithm training phase are the data originally collected from the power station including the weather forecast for these days as we received them from the tomorrow.io API. Since our total data from the power station cover a 6 month period, part of this data will be used as our training set and the rest will be used for the evaluation of the algorithm performance. Table I shows the features used to train the random forest machine learning model with the active power being the station’s power production. The same vector of data (minus the active power) is provided to the trained model in order to get the predicted power output values.

Using the data from our dataset, we then train and test the Random Forest algorithm, and validate our data over the actual values to get the following results. We feed into our model a total of 8 parameters. The time of the day, the solar irradiance, the outdoor temperature and humidity, the wind speed and direction as it affects the performance of the PV panels. We also provide our model with information about the weather conditions (cloud coverage and weather code) as cloudy or rainy days severely reduce the output of the power plant and have a huge effect on the output of the system. The most influencing point in the input vector is in fact the irradiance provided by the weather forecast, with minor contributions from the temperature, and wind conditions, as seen in Table II. The limited effect of the weather conditions in the total forecast can be at some point attributed to the period covered by the dataset, as it contains mainly sunny days during summer in Greece. This is something we expect to change in the next months as our dataset increases. Our implementation uses a Random Forest Regressor with 1000 estimators and 70% of the dataset as a training dataset, and with the rest (30%) as test data. The estimation managed a coefficient of determination of 0.93 and an RMSE value of 363.

The average power production recorded with the on-site sensors and the average forecasted power production for every hour of the day is presented in Figure 4. In the same figure we also present the mean root square error for every hour of the day. Unlike the model based case, we observe here that the root mean square error follows the normal distribution during the day, with higher values around noon as the irradiance forecast had a higher deviation during those hours. We can see here how the model trained using real data from the PV power station is capable of calculating more precisely the expected power production. This of course is possible after the collection of data from the operation of the power station for a period of at least 6 months, during which we observed multiple weather conditions and their effects on the station’s outputed energy. For the first days and months of the operation in a new power station this solution would not be able to provide us with such accurate estimations, and thus we will need to operate using the data from a mathematical model as the one presented in the previous section.

![Fig. 4. Average Power Production and Average Forecasted Power using the machine learning based power forecasting along with the forecast's root mean square error for every hour of the day](image)

For our analysis we used Random Forest algorithm instead of other solutions as the nature of the data we wanted to predict were significantly dependent on the values of the solar irradiance, and the environmental conditions occuring on site (e.g., sunny or cloudy days, rain or high winds/temperatures). Other techniques and algorithms like ARIMA, SARIMA [13] and Prophet [14] use only historical data and seasonal information to predict the trends of the value we are trying to forecast. This is not adequate for us, as we cannot have historical data for energy production at any location, and even if we did, our predictions would not be taking into account short term weather conditions like cloudy or rainy days. If we want to estimate the total power production over larger periods (e.g., over weeks or months) the effects of weather phenomena can be eliminated and such algorithms could be used.

C. Operational Insights

Using the data computed above, we were able to come back to the recorded data for each day and compare them with our predictions. This comparison helped us understanding the days
and the time intervals during which the power station underperformed for some reason, according to the existing environmental conditions. This under-performance can be transient and disapper after a couple of hours or extend over multiple hours and possible days, indicating a malfunction, error, or severe problem that needs immediate attention from the owner or the support crew of the power station.

We applied this logic in the whole period of our dataset in an attempt to identify such days that would prove our theory correct. We in fact detected two periods in August 2021, where the power station severely under-performed when compared to the expected power production. The two periods, on the 6th and 22nd of August, spanned only a few hours each day but happened during the middle of the day, hours during which the power station provides the network with the most power throughout the day. For both days, we were able to confirm that the lower production was in fact due to an automated system shutdown, initiated for safety reasons by the ABB REF615 automated feeder protection and control relay installed on-site. This device, provides us with a sensor element that is read through the Modbus interface discussed in Section IV called `circuitBreakerOpen` that shows when the main system’s circuit breaker is open, resulting in no power provided to the external grid. This breaker is disconnected automatically when the feeder relay detects an instability in the energy produced and is reconnected either automatically if safe and possible, or manually if the problem that caused the disconnection needs human intervention and system inspection. In our two detected cases, the event was transient and the system recovered automatically.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

This paper presented results in two different fronts. On the first front, we showcased a system that is capable of collecting data from PV systems and power stations using edge computing and cloud services. On the second front, we presented our work on forecasting power production from PV systems using physical models and machine learning approaches. We noted the qualities of each method and compared their results with the original data. Finally, we used the data from the power production forecasts to benchmark the operation and the performance of the PV power station and correctly identified days where the power production was significantly reduced from the forecasted expectations. As our next steps, we plan to expand our test sites to new PV power plants, and work on detecting more conditions inside the PV station based on its power production.

REFERENCES