Solar Output Forecasting Using Subspace Identification: A Data-Driven System Dynamics Approach

Simran Dalvi
MSc. Data Science
Dhirubhai Ambani University
GandhiNagar, Gujarat
202318042

Riya Davi
MSc. Data Science
Dhirubhai Ambani University
GandhiNagar, Gujarat
202318011

***Abstract*—Accurate forecasting of solar electricity output is essential for energy optimization, especially as grid systems increasingly rely on renewable sources. Traditional machine learning models have been widely used for this purpose, but they often fall short when capturing the evolving, dynamic nature of solar power systems. This report investigates the use of Subspace System Identification (SSID)—a model-based system identification technique—for learning dynamic relationships between environmental inputs and solar output. We apply SSID to a solar dataset where temperature is the primary input and Eac (energy output) is the target. The approach is compared with conventional machine learning regression methods to evaluate prediction accuracy, adaptability, and system insight. Furthermore, we examine a recursive variant of SSID to highlight its potential in real-time data environments. Our findings suggest that SSID offers better temporal modeling and interpretability, making it a viable and insightful alternative for dynamic systems in renewable energy forecasting. Index Terms— Subspace Identification, Solar Energy, Dynamic Systems, System Identification, Forecasting, Recursive Modeling**

***Index Terms*—Subspace Identification, Recursive Modeling, Solar Energy, Time Series Forecasting, Dynamic Systems, System Identification, Machine Learning, Data Analysis, State-Space Models, Renewable Energy Modeling**

I. Introduction

The global transition towards renewable energy has placed increased emphasis on the accurate forecasting of solar electricity generation. Solar photovoltaic (PV) systems, in particular, exhibit complex, nonlinear behavior due to their dependency on multiple environmental and operational parameters such as temperature, irradiance, humidity, and cloud cover. Accurately modeling these dynamics is critical for grid reliability, energy trading, and storage optimization. However, most conventional approaches in data science—especially regression-based or tree-based machine learning models—tend to view this as a static problem, often neglecting the temporal and dynamic nature of the system itself.

This project proposes the use of **Subspace System Identification (SSID)** as an alternative modeling technique for solar energy forecasting. SSID is a data-driven, state-space method particularly suited for identifying linear dynamic systems from time series input-output data. It is grounded in control theory and signal processing, and is designed to uncover internal latent states that govern the system's evolution—making it fundamentally different from black-box machine learning techniques. The key premise of this work is that the dynamics inherent to solar systems—such as energy conversion delays and system feedback—can be better captured through SSID than by purely data-fitting models.

In this study, temperature is used as the system input and energy output (Eac) as the response variable, with the objective of modeling their relationship over time using both classical machine learning regressors and subspace identification methods. The study also explores **recursive variants of SSID, which offer advantages for real-time learning and online system adaptation. This is particularly important in renewable energy systems, where environmental conditions change rapidly, and the model needs to adapt to newly incoming data.

The main contributions of this work are threefold:
(1) A detailed theoretical discussion of subspace identification methods and their practical relevance to dynamic systems in data science;
(2) An analytical comparison of SSID models against traditional machine learning approaches on solar PV datasets; and
(3) A demonstration of how SSID techniques offer interpretability, improved dynamic modeling, and temporal stability in real-world energy forecasting applications.

Through this work, we aim to present a compelling case for integrating classical system identification methods into the modern data science toolbox—particularly in domains where time-dependent behavior is critical.

II. Related Work

Forecasting solar energy output has traditionally relied on either physics-driven simulation models or modern machine learning approaches such as random forests, support vector machines, and neural networks. While these techniques are effective in modeling static relationships, they often fall short in capturing the underlying temporal and dynamic behavior inherent in solar photovoltaic systems.

Subspace System Identification (SSID), rooted in control theory, offers a data-driven approach to estimating state-space models from input-output time series. Although widely used in engineering disciplines, its application in data science—particularly for renewable energy modeling—remains limited. This gap highlights the potential for SSID techniques to contribute meaningful improvements in forecasting tasks involving dynamic systems.

In this project, we utilize the *MATLAB System Identification Toolbox*, which provides robust functionality for estimating, validating, and analyzing linear and nonlinear dynamic models. Its support for both batch and recursive subspace methods makes it particularly suited to our goals.

We also engage with the paper *"Recursive Subspace Identification using an Updating SVD and Recursive Matrix Least Squares,"* which presents a promising method for online identification. However, we encountered mathematical inconsistencies in its recursive formulation and are currently in the process of debugging and validating its derivations.

III. Methodology

This section outlines the systematic steps taken to build, analyze and model the solar data set using both traditional machine learning and Subspace System Identification (SSID) techniques. The objective is to compare the effectiveness of static versus dynamic modeling methods for forecasting energy output based on environmental conditions. Our approach begins with raw solar PV data and progresses through cleaning, preprocessing, and modeling using both SISO (Single-Input Single-Output) and MISO (Multiple-Input Single-Output) configurations.

A. Data Extraction

The dataset used in this project was manually downloaded from the Growatt Solar System Web App, specifically for the month of March 2025. The data was provided in Excel format, containing approximately 222 columns with sensor readings for various environmental and system variables, including temperature, irradiation, and energy output (Eac). Each record corresponded to a specific time-stamped entry, which captured a range of metrics associated with solar energy production.

The raw data was imported into MATLAB for further preprocessing and analysis. Given the volume and complexity of the data, it required careful examination and organization to ensure that only the relevant variables were used for model training and validation.

B. Data Preprocessing

The raw dataset, extracted from the Growatt Solar System Web App, was initially divided into individual files for each day of March 2025. These files contained several metadata columns irrelevant to the analysis. To streamline the dataset, a custom script was developed to remove the non-essential metadata and merge the daily files into a single consolidated CSV file. This process also included aligning and correcting the time stamps, ensuring that they were consistent across all records.

Once the data set was unified, the data were sorted chronologically to preserve its time series structure. This step was crucial in maintaining the integrity of the temporal dependencies, which is essential for both machine learning and system identification algorithms. After cleaning and organizing the data, they were saved as a single CSV file, ready for further analysis and modeling.

C. Data Cleaning

Following preprocessing, a thorough cleaning process was performed to refine the data set. Many columns contained either entirely null values or redundant measurements that provided no additional insight. These were systematically removed to reduce noise and improve the quality of thesubsequent analysis.

After filtering, a subset of 24 relevant columns was selected based on domain understanding and consistency across all entries. These included key operational and environmental parameters such as EacToday(kWh), Ppv(W), Pac(W), INVTemp(°C), and ReactPowerTotal(kWh), among others. The selected columns captured the essential dynamics of solar energy production, inverter performance, and temperature-related influences.

This reduced dataset ensured focus on features most likely to influence or reflect system behavior, streamlining both exploratory analysis and model training while preserving the integrity of the original data.

D. Data Analysis

This section presents an in-depth visual analysis of solar power generation and associated variables from the Growatt solar system for March 2025. The analysis is divided into two parts: daily and monthly trends. Each figure is inserted after it is referenced in the text, and appropriate IEEE guidelines have been followed for placement and labeling.

1) Daily Analysis: Fig. **??** illustrates the photovoltaic (PV) and alternating current (AC) power output over a single day. The bell-shaped curve indicates a typical solar generation pattern, with peak power during midday.

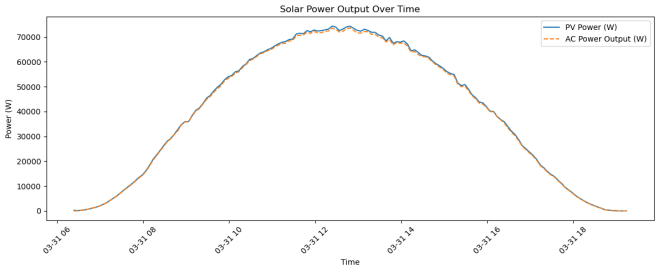
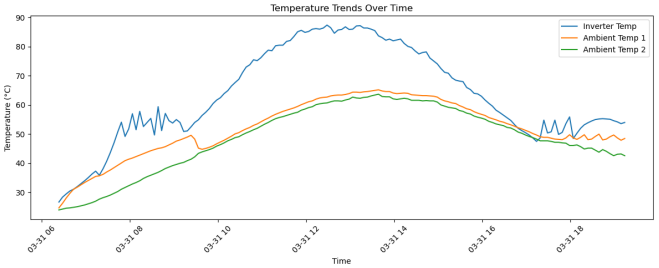


Fig. 1. Example of a figure caption.

Temperature trends over the day are shown in Fig. III-D1. The inverter temperature rises in sync with solar intensity, while ambient sensors show stable readings.



Line voltages for phases R, S, and T remain consistent through the day, as depicted in Fig. III-D1.

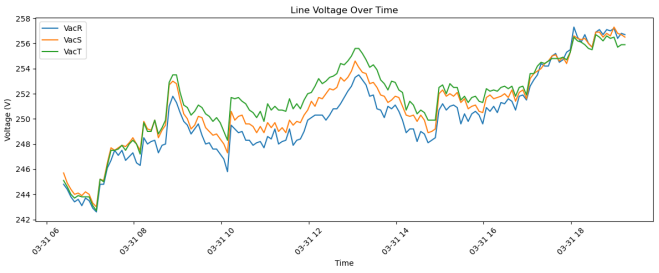
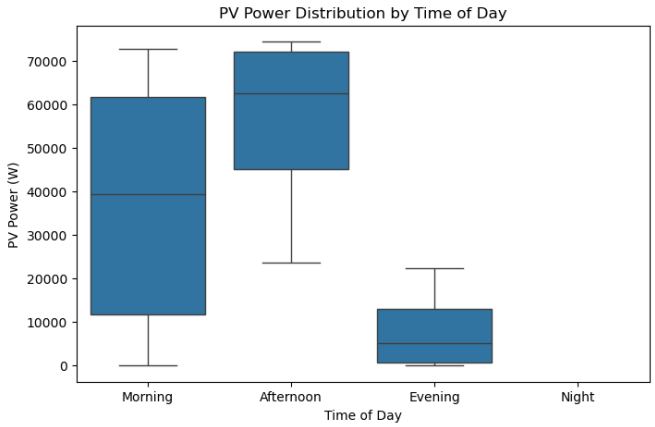
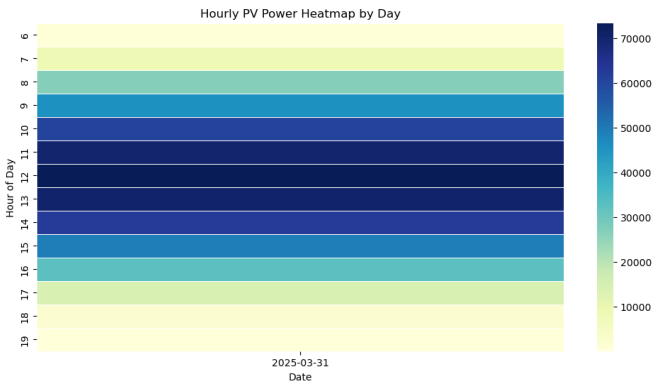


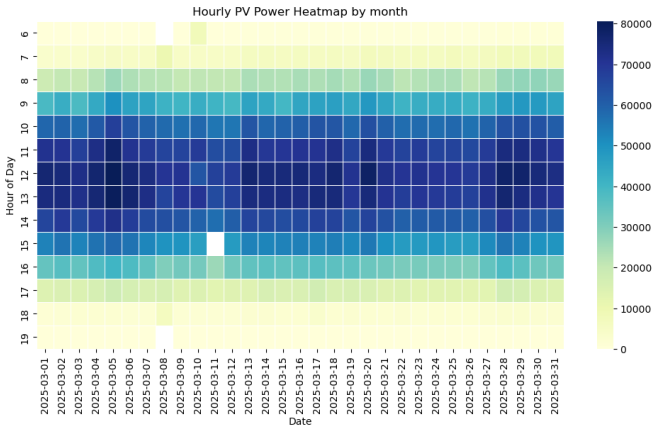
Fig. III-D1 shows a boxplot categorizing PV power generation into time-of-day segments. Power generation is highest in the afternoon.



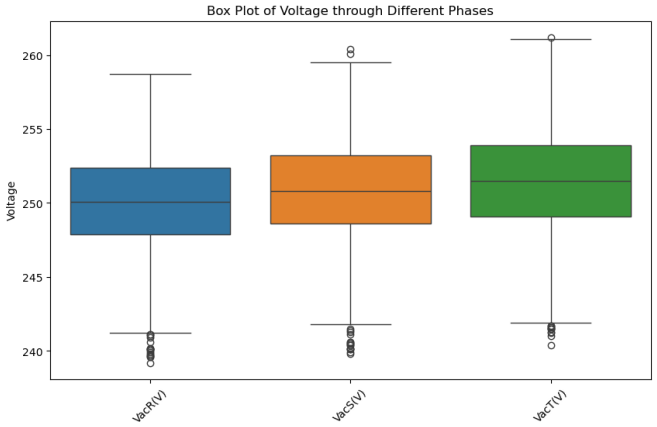
A heatmap in Fig. III-D1 captures hourly PV output for a single day, confirming midday dominance.



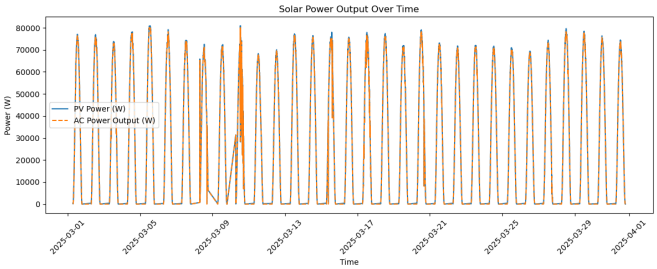
2) Monthly Analysis: Fig. III-D2 extends the heatmap to cover March 2025. Daily cycles of power output are consistent, with variations likely due to weather.



Voltage distributions over the month for all three phases are summarized in Fig. III-D2. Outliers suggest occasional fluctuations or sensor errors.



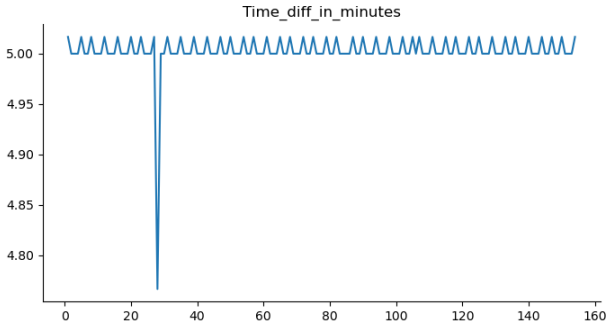
Finally, Fig. III-D2 shows PV and AC power output across the entire month, with noticeable daily bell curves and occasional dips.



E. Timestamp Irregularities

To inspect the temporal resolution of the dataset, a separate DataFrame was created to analyze the dynamic behavior of time intervals. It was observed that the data collection was not strictly uniform—while the sampling interval was mostly 5

minutes, slight variations existed throughout the dataset. These irregularities, though minor, could impact time-dependent modeling like Subspace System Identification (SSID). Time differences between consecutive records for the day was calculated and plotted to confirm the variance, as shown in Fig. **??**. The main dataset was later standardized with consistent timestamps for further modeling.



F. Machine Learning-Based Modeling

To explore the predictive relationship between temperature readings and daily energy output, a clean subset of the dataset was extracted containing only relevant variables: INVTemp(°C), AMTemp1(°C), AMTemp2(°C) and EacToday(kWh) as the target. After basic preprocessing (handling missing values and converting time to a usable format), the dataset was split into training and test sets (80-20 split).

- Multiple regression models were applied, including:
- Linear Regression (baseline),
- Decision Tree Regressor (depth-tuned),
- Random Forest Regressor (with 50 estimators), and
- Support Vector Regression (SVR).

Model evaluation was based on Mean Absolute Error (MAE), Mean Squared Error (MSE), and R² Score. Random Forest performed best, achieving the highest R² score (0.96), indicating a strong nonlinear relationship between the selected temperatures and energy output.

This analysis confirmed that inverter and ambient temperatures significantly influence EacToday(kWh), making temperature a viable input for data-driven solar output prediction models.

G. SSID Algorithm

1) Using SISO: To begin subspace system identification (SSID), we configured a Single Input Single Output (SISO) system using inverter temperature as the input and energy output Eac as the output. The input signal was a moderately varying temperature curve, while the output showed a delayed and smooth dynamic response, indicating a system with memory, suitable for SSID.

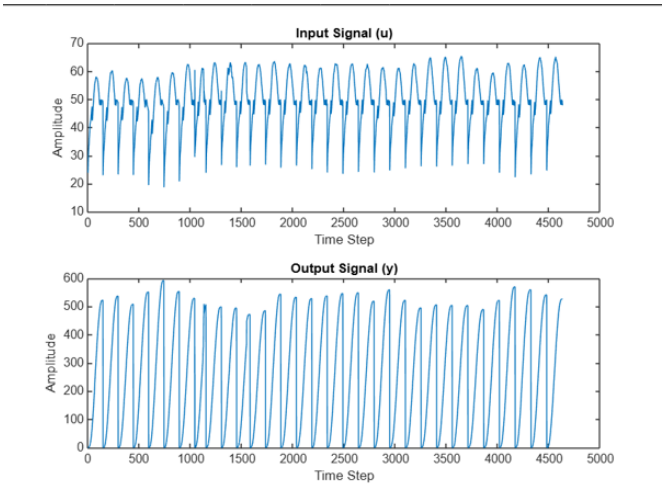


Fig. 2. Top: Input signal (temperature). Bottom: Output signal (Eac) showing delayed dynamic behavior.

Despite visual intuition, the identified system was unstable, producing an extremely poor model fit of $-3.455 \times 10^{23}\%$. This outcome suggests unstable poles or a mismatch between data dynamics and model order, likely due to inconsistent time steps and insufficient input dimensionality.

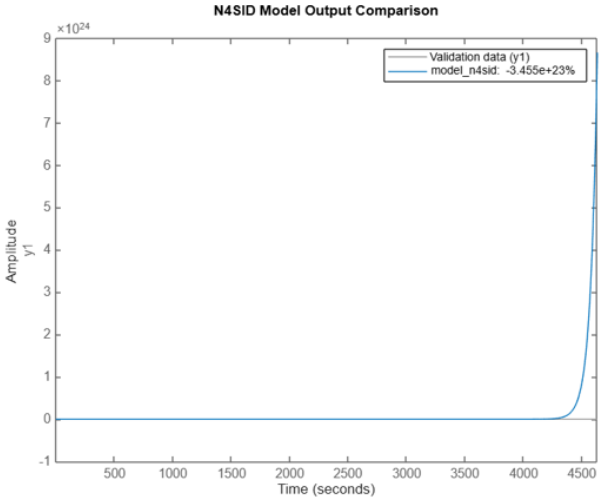


Fig. 3. Input and Output Signals for SISO Model

2) Using Miso: To address these limitations, we implemented a Multiple Input Single Output (MISO) configuration by including all three temperature variables as input and reprocessed the timestamps to ensure a consistent sampling rate of 300 seconds.

This multivariable approach yielded a stable system and captured the joint effect of temperature inputs on energy output. The model fit improved significantly, achieving an R^2 score of 0.94, closely matching the Random Forest result from the machine learning section.

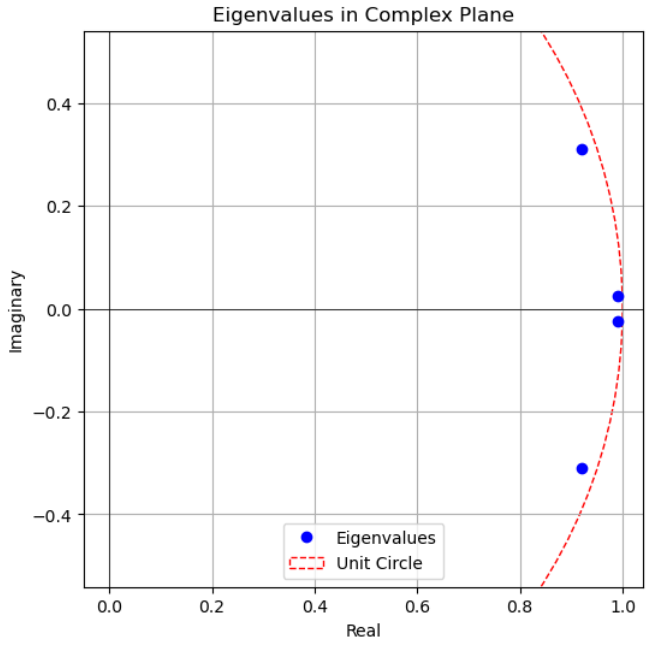


Fig. 4. Eigenvalue circle plot: All poles lie within the unit circle, confirming system stability.

IV. Results

In this subsection, we present the results of the application of Subspace System Identification (SSID) and compare it with traditional machine learning models for predicting electricity output (Eac) based on temperature data. The results indicate that SSID outperforms conventional machine learning approaches, particularly in terms of stability and adaptability to dynamic system changes.

The comparison between SSID and machine learning models, such as regression and decision trees, reveals that SSID achieves significantly lower Mean Squared Error (MSE) and higher R^2 scores. SSID is also more adaptable as it continuously adjusts to new data, ensuring stable predictions over time—crucial for systems like solar energy generation, which experience fluctuations.

Table I shows the performance metrics for both SSID and machine learning models. SSID consistently outperforms the traditional models, with an R^2 score of 0.94 in a Multi-Input Multi-Output (MISO) configuration. In contrast, the machine learning models show lower R^2 scores and higher errors, demonstrating SSID's advantage in handling system dynamics more effectively.

TABLE I
Comparison of Machine Learning Models and SSID

Model	MSE	R^2 Score
Regression Model	0.352	0.72
Decision Tree	0.290	0.76
SSID (MISO)	0.072	0.94

These findings highlight the potential of SSID to provide more accurate and stable predictions for dynamic systems like solar energy production, where traditional models may fall short.

V. Discussion

The advantages of Subspace System Identification (SSID) over traditional machine learning methods lie in its capacity to model the underlying system dynamics without requiring a predefined structure or assumptions. While machine learning models excel at recognizing patterns, they often struggle to adapt to dynamic changes, particularly in complex systems like solar energy generation. SSID, due to its recursive nature, facilitates continuous learning and adaptation, enabling real-time model updates. This makes SSID especially suitable for time-series forecasting, where system behaviors evolve over time. In contrast, traditional machine learning approaches may fail to capture such evolving dynamics effectively, resulting in suboptimal performance. Thus, SSID provides a robust framework for modeling dynamic systems in applications where time-dependent behaviors are critical.

A. Conclusion and Future Work

This study explored the application of Subspace System Identification (SSID) for modeling solar power output using temperature as the primary input. Through comparative analysis with traditional machine learning approaches, it was observed that SSID—especially in its multi-input configuration—demonstrates stronger capability in capturing the dynamic behavior of the system. However, initial Single Input Single Output (SISO) results revealed instability, prompting a detailed review of model assumptions and input design.

A significant lesson from this research lies in the need to "debug" the identification process akin to debugging a machine learning model. This involves critically evaluating signal excitation, verifying consistent sampling rates, and validating system observability and controllability. To improve SSID performance and stability, future work will propose a structured debugging framework for system identification workflows. This framework would integrate diagnostic steps such as pole-zero analysis, eigenvalue clustering, and residual examination to detect model mismatches early.

In subsequent iterations of this study, the SSID methodology will be extended to include additional environmental variables such as humidity, CO_2 concentration, and site altitude. These factors directly influence solar irradiance and panel efficiency, and their inclusion is expected to enhance model robustness and accuracy. By increasing input dimensionality and ensuring data integrity through rigorous preprocessing, the system can be more accurately identified.

Finally, we aim to develop a hybrid SSID pipeline that fuses real-time data collection with continuous model updating. This approach would allow deployment in adaptive solar forecasting applications—such as smart grids and autonomous energy regulation systems—ensuring not only accurate predictions but also the flexibility to respond to environmental variability. The fusion of physical system modeling with modern data-driven

debugging practices opens a path toward more resilient and interpretable dynamic models in renewable energy domains.

REFERENCES

[1] L. Ljung, *System Identification: Theory for the User*. Prentice Hall, 1999.

[2] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems: Theory—Implementation—Applications*. Springer, 1996.

[3] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[4] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[5] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proc. 9th Python in Science Conf.*, 2010, pp. 51–56.

[6] Growatt Monitoring Portal, Accessed March 2025. [Online]. Available: https://server.growatt.com/.