Applied Analytics Practicum

Enverus

Team 5

hdupouy3@gatech.edu  
oalmansour3@gatech.edu  
[jgarcia340@gatech.edu](mailto:jgarcia340@gatech.edu)

***Abstract—*** This project presents a study conducted to evaluate the impact of various locations and modeling techniques on the prediction of solar farm performance. The research project was undertaken in collaboration with Enverus, an energy technology company renowned for its innovative software, data, and services in the energy market. The objective of the study is to compare and contrast different design options by incorporating location-specific source data and alternative modeling techniques. The paper discusses the problem statement, objectives, and desired goals of the investigation.

# Introduction

## Enverus

## Enverus, an established energy technology company, is at the forefront of this research project. With its inception in 1999, Enverus has become a prominent provider of energy market data, analytics, and technology solutions (Enverus, 2023). The company's core objective is to optimize operations and foster a deeper understanding of energy markets through innovative software, data, and services. Enverus empowers energy firms with platforms, tools, and applications to navigate the complex and ever-changing market landscape. Moreover, Enverus offers a range of invaluable services, including expert guidance, data analysis, and market intelligence, solidifying its position as a trusted leader in the industry.

## Objective

### Project objectives

In this project, our primary objective is to study the influence on the performance prediction of solar farms that may be achieved by adding location-specific source data and various modeling methodologies. We will evaluate and contrast several design possibilities in order to give insights into improving the operations of solar farms and the decision-making processes involved. This will be accomplished by undertaking an extensive study and testing. Through working together with Enverus and making use of data sets that have been anonymized (data sets encompassing input and target variables based on the data from Californio ISO toward the actual solar mega-watt generation (California ISO, 2023) , we will be able to successfully accomplish the study goals that we have set for ourselves. The results of this research will add to the current literature on the prediction of the performance of solar farms, and they will also aid energy firms like Enverus in making educated choices so that they may prosper in the ever-changing energy market.

### Vale and Magnitude

### Value -This project holds substantial value for the company by generating valuable insights and knowledge concerning the prediction of solar farm performance. Through the integration of location-specific source data and advanced modeling techniques, the project aims to enhance the accuracy and predictive performance of solar farm performance models. By improving the precision and reliability of these models, energy companies can make more informed decisions regarding their operations, resource allocation, and identification of regions with high solar energy generation potential. This optimization of operations and resource utilization can lead to cost savings and overall performance improvements, enabling companies to stay competitive in a rapidly changing market.

### Magnitude - The magnitude of this project is noteworthy due to its comprehensive approach and incorporation of multiple elements. The project involves evaluating various design options, including diverse data sources and model classes, to identify the most effective combination for solar farm performance prediction. It requires access to anonymized data sets containing input and target variables and leverages open-source machine learning and visualization libraries for analysis. The scale of the project involves analyzing a substantial amount of data, with up to 50,000 rows and 20 input variables per data source. Furthermore, advanced modeling techniques, such as Neural Networks and Tree Models, are utilized to capture complex relationships and patterns within the data. The project also includes a thorough review of relevant literature on solar generation forecasting to incorporate best practices and industry knowledge. Given the comprehensive nature of the project and its potential impact on decision-making processes within energy companies, the magnitude of this endeavor is significant.

## Problem Statement

Existing methods for forecasting solar farm performance have limitations due to a lack of consideration for location-specific data and reliance on traditional modeling tools. This leads to imprecise and unreliable predictions. To address this issue and improve the accuracy and foresight of solar farm performance models, it is necessary to evaluate and compare the outcomes obtained by utilizing location-specific data sources and advanced modeling methodologies. By doing so, energy providers can enhance operational efficiency, optimize resource allocation, and identify areas with the highest potential for solar power production.

Research questions:

• When comparing generic data sources and traditional modeling methods to solar energy-generating performance prediction, how does integrating location-specific data sources and sophisticated modeling techniques affect the accuracy and predictive performance?

• Moreover, how might these enhancements help energy firms maximize operations, allocate resources, and pinpoint areas with the most significant potential for solar power generation?

## Hypothesis

The combination of location-specific source data and advanced modeling techniques will considerably enhance the accuracy and predictive performance of solar farm performance prediction when compared to the use of generic data sources and conventional modeling approaches.

"The hypothesis is founded on the premise that the performance of solar farms is influenced by a variety of factors, including geographical location and modeling technique. It is anticipated that by utilizing location-specific source data, which takes into account the spatial proximity to the region under study, the models will capture the unique characteristics and variations associated with different locations. This data may include information about solar irradiance, weather patterns, terrain features, or the proximity of other environmental factors.

This study's design matrix incorporates numerous data sources that represent macro, meso, and microregions in close proximity to the target area. In addition, numerous time-series models, including Linear Regression, Random Forest, and XGBoost, are incorporated into the endeavor. This project's primary objective is to answer the research questions and gain a comprehensive understanding of the factors influencing the performance of solar farms.

In addition, the hypothesis assumes that advanced modeling techniques, such as Neural Networks, Tree Models, and other sophisticated algorithms, will enable the models to capture complex relationships and nonlinearities in the data. These techniques have the potential to identify patterns and correlations that conventional linear regression models may not adequately capture.The hypothesis suggests that by integrating sophisticated modeling techniques with location-specific data sources, the models will be more accurate and provide more reliable predictions of solar farm performance. This, in turn, can aid energy companies in optimizing their operations, allocating resources, and identifying regions with the greatest potential for solar energy generation.

To validate this hypothesis, the project will assess the performance of various design options, such as diverse data sources and model classes, in order to evaluate their efficacy. The evaluation of the models' accuracy, robustness, and generalizability will shed light on the efficacy of the proposed approach and support the hypothesis if significant improvements are observed.

The anticipated findings of this study hold the potential to provide valuable insights into the renewable energy industry. By evaluating various modeling techniques and analyzing the impact of location-specific data on solar farm performance, this project aims to enhance the accuracy and efficiency of predictive models. The outcomes of this research can contribute to the development of more reliable and effective modeling techniques, enabling better decision-making in the planning and implementation of solar farms.

Ultimately, this project seeks to promote the adoption of sustainable energy solutions and drive the advancement of the renewable energy sector by providing valuable knowledge and facilitating informed decision-making processes.

# literature review

## Review

The geographical location within California is a crucial factor influencing the efficacy of solar farms, according to our California-specific study. Miao, Ning, Gu, Yan, and Ma (2018) highlight the variability of solar radiation across different locations and propose a comprehensive framework that takes latitude, longitude, and weather patterns into account to evaluate their impact on energy production. Miao, Ning, Gu, Yan, and Ma's (2018) research emphasizes significant differences in performance based on specific geographical locations, highlighting the significance of accurate modeling techniques. Consequently, the objective of our project is to pinpoint a specific location within California for robust data analysis and the application of advanced data science techniques for predicting time series data.

According to Singh and Singh (2021), solar energy production is affected by a variety of environmental factors. These include solar irradiance, temperature, humidity, pollution, shading, and wind velocity. In addition, the technical design characteristics of photovoltaic cells, such as the materials employed, have an effect on their ability to generate electricity (Chikate & Sadawarte, 2015). Therefore, we will acquire these additional data points during our data collection procedure in order to obtain a comprehensive understanding of the significant factors influencing the performance of solar energy generation.

Similar studies in the field of solar energy generation emphasize the need for reliable modeling techniques on a consistent basis. Hobbs et al. (2022) conducted research on probabilistic solar prediction utilizing the probabilistic Watt-Sun model, emphasizing the need for reliable and accurate models despite the inherent predictability uncertainties. Similar to this, Aksoy and Genc (2023) proposed an ensemble model that incorporates various fundamental models, such as Random Forests and Gradient Boosting Machines, to improve the precision of energy production forecasts. Aksoy and Genc's research demonstrates the efficacy of ensemble learning in capturing the complex relationships between input variables and energy production (2023).

These studies contribute to the existing knowledge base in the field of solar farm performance prediction and highlight the importance of accurate modeling techniques and consideration of site-specific factors. By incorporating these findings into our research, we hope to improve the precision and dependability of our solar energy generation performance prediction models.

# methodology

## ETL

Figure 1: Extract, Transform, and Load (ETL) Data Pipeline

The methodology employed in this study involves the implementation of an Extract, Transform, and Load (ETL) data pipeline, as depicted in Figure 1. The ETL process is crucial for collecting, preparing, and organizing the data to ensure its suitability for analysis and modeling.

1.Extract: In the extraction phase, we obtain the necessary data from multiple sources, including the CAISO Total Solar Historical Actuals dataset provided by the project owners ( see appendix 1). This dataset contains the date, hour, and actual MW metrics data for solar farm performance. Additional data sources may include geographical information, weather patterns, solar irradiance data, and other relevant variables.

2. Transform: Once the data is extracted, we perform various transformations to clean, preprocess, and enhance the dataset. This involves steps such as handling missing values, removing duplicates, standardizing data formats, and integrating additional data sources. We also conduct feature engineering to create new variables or derive meaningful insights from the existing data.

3. Load: The transformed data is then loaded into a suitable storage system or database for further analysis. This enables easy access and retrieval of the data during the modeling and evaluation stages. The loaded data is structured and organized in a way that facilitates efficient data manipulation and modeling tasks.

The ETL data pipeline ensures that the data used in our study is of high quality, consistent, and ready for analysis. By employing this robust methodology, we can effectively process and prepare the data for subsequent modeling and evaluation processes.

The next section will elaborate on the specific modeling techniques employed in this study to predict solar farm performance.

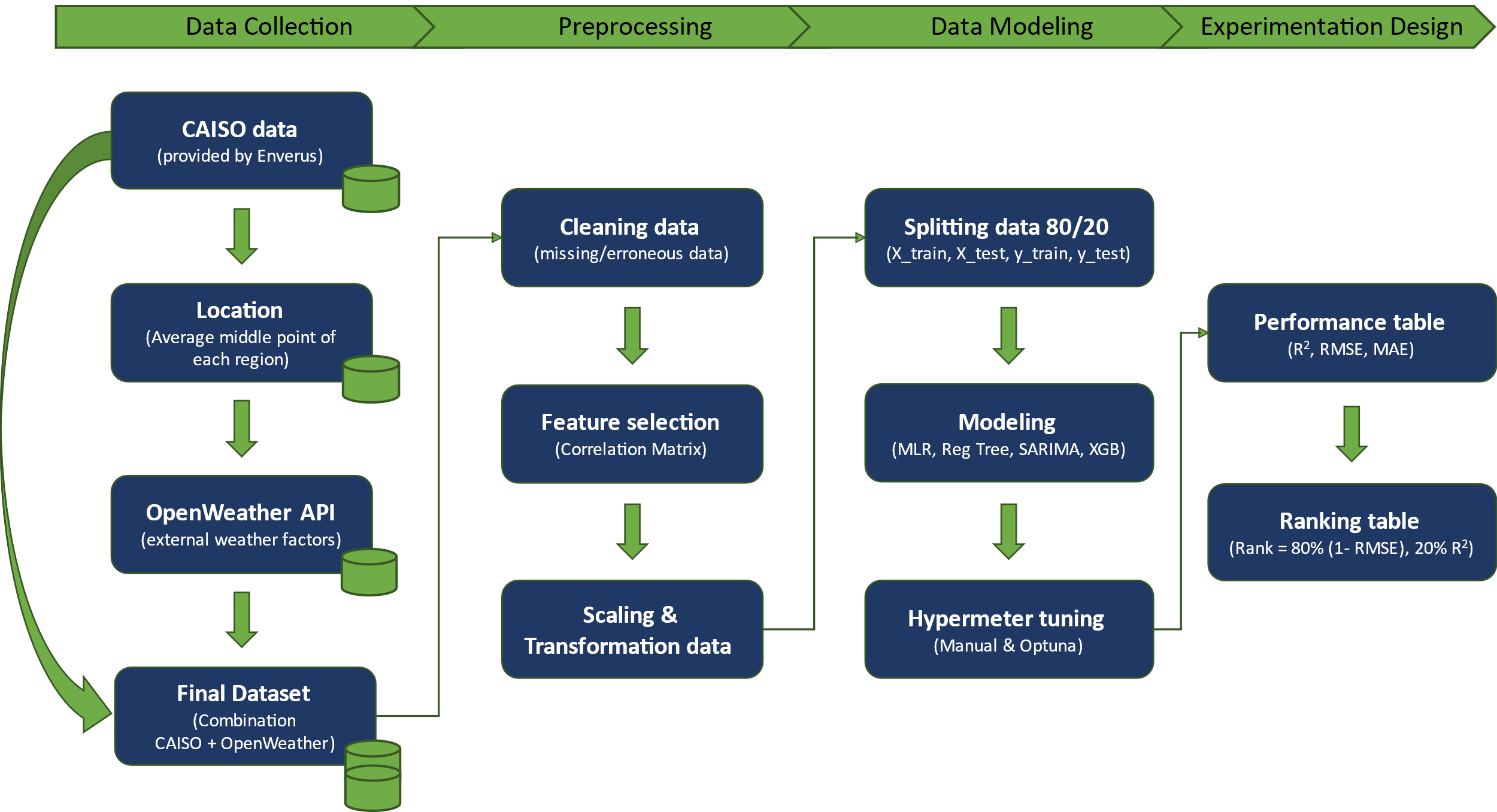


Figure 1. ETL Data Pipeline

## Data Collection

The data collection process for this project is designed to gather relevant information for assessing the impact of geographical locations and modeling techniques on the prediction of solar energy generation performance. The following steps outline the data collection approach:

1. Identify Data Sources: Firstly, we will identify the necessary data sources to support our analysis. As part of our collaboration with Enverus, we have access to simplified datasets containing timestamps and the actual measure of mega-watts generated. However, since this measure represents an average location of the three main regions in California, we need to determine a specific location to retrieve future data. To obtain environmental factors such as temperature, humidity, dust, shading, and wind speed, we will utilize data from OpenWeather, a reliable source of weather information (OpenWeather, 2023) (see appendix 2).

2. Obtain Data: Once the data sources are identified, we will gather the necessary data from these sources. This includes retrieving Enverus proprietary datasets, accessing publicly available datasets, exploring academic research repositories, and reviewing industry reports. The collected data will be acquired in a structured format to facilitate efficient preprocessing and modeling in later stages.

3. Ensure Data Quality: To maintain a high level of data quality, we will implement quality assurance measures. This involves verifying the reliability and integrity of the data sources, conducting checks for missing or erroneous data points, and addressing any data inconsistencies or outliers. If required, data cleansing techniques will be applied to enhance the overall quality and reliability of the dataset.

By following this data collection process, we aim to obtain a comprehensive and reliable dataset that encompasses relevant variables for analyzing solar farm performance. The gathered data will serve as the foundation for subsequent preprocessing, modeling, and evaluation stages of the project.

## Preprocessing

## Data preprocessing plays a crucial role in preparing the collected data for analysis and modeling in our study. The following steps outline the preprocessing techniques applied to the dataset:

## 1.- Data Cleaning: We begin by addressing any missing, erroneous, or inconsistent data points identified during the data collection phase. Missing values can be imputed using appropriate techniques, such as mean imputation or regression imputation. Erroneous or inconsistent data points are corrected or removed based on predefined criteria. This ensures the integrity and quality of the dataset.

## 2.- Feature Selection: Next, we analyze the collected dataset to identify the most relevant features that significantly contribute to the prediction of solar farm performance. This step involves establishing a correlation matrix between variables, performing feature importance analysis using techniques like information gain or recursive feature elimination, and utilizing domain knowledge to select the most informative features for modeling. By selecting the appropriate subset of features, we can reduce dimensionality and improve the efficiency and interpretability of the models.

## 3.- Feature Scaling and Transformation: To ensure compatibility and optimal performance of machine learning algorithms, certain features may require scaling or transformation. Numerical features are typically normalized to a common scale, such as using techniques like min-max scaling or standardization. Categorical features are transformed into dummy variables or one-hot encoded to represent them numerically. This allows the models to interpret all features as numerical inputs.

## 4.- By applying these preprocessing techniques, we aim to prepare the dataset in a clean, concise, and standardized format, suitable for further analysis and modeling. This ensures that the data is appropriately processed and ready for the subsequent stages of the study, such as model development and evaluation.

## Data Modeling

The data modeling phase is centered around developing models that can effectively predict solar farm performance based on the collected and preprocessed data. The following steps outline the approach taken in this phase:

1.- Model Selection: Building upon the objectives of the study and the insights gained from the literature review, we will select a range of modeling techniques to explore. This will include traditional statistical models such as multi-linear regression, as well as machine learning algorithms like random forests and extreme gradient boosting. The selection of these models allows us to leverage both established and cutting-edge approaches in predicting solar farm performance.

2.- Model Development: In this step, we will develop and train the selected models using the preprocessed dataset. The dataset will be split into training and test subsets, ensuring that the models are trained on a representative portion of the data and evaluated on unseen data. Model parameters will be optimized using techniques such as grid search or random search to find the best configuration for each model. This iterative process will involve fine-tuning the models to maximize their performance.

3.- Model Evaluation: The developed models will undergo a comprehensive evaluation process. We will utilize cross-validation techniques to assess the models' generalization performance and robustness. This involves dividing the data into multiple subsets, training the models on one subset, and evaluating their performance on the remaining subsets. Evaluation metrics such as root mean squared error (RMSE), R-squared, and mean absolute error (MAE) will be used to measure the models' predictive capabilities. Additionally, considerations will be given to factors such as accuracy, precision, recall, and computational efficiency to ensure a well-rounded evaluation.

By following these steps, we aim to develop accurate and reliable models that can effectively predict solar farm performance. The iterative nature of model development and evaluation allows for continuous improvement and refinement, ensuring that the final models are robust and capable of making informed predictions.

## Experimentation Design

The experimentation design phase is a critical component of the project, as it allows us to compare and contrast different design options, considering both geographical implications and modeling techniques. The following steps outline the approach taken in this phase:

1. Design Matrix: We will construct a design matrix that includes data combinations from the California CAISO and OpenWeather datasets. This design matrix will encompass macro, meso, and micro regions in proximity to the solar farm location. By incorporating a variety of geographical data sources, we aim to capture the spatial variations and unique characteristics of different locations.

2. Dataset Split: The design matrix will be used to split the dataset into training and testing subsets. Both the response and dependent variables for our models will be divided into an 80% training dataset and a 20% testing dataset. This ensures that the models are trained on a representative portion of the data and evaluated on unseen data.

3. Model Selection and Setup: The modeling phase will involve the implementation of various model classes, including multi-linear regression, regression tree, Seasonal Autoregressive Integrated Moving Average (SARIMA), and eXtreme Gradient Boosting (XGBoost). Each model class will undergo an experimental setup phase, where different parameters will be tested and optimized to enhance their performance. This iterative process allows us to fine-tune the models and maximize their predictive capabilities.

4. Performance Comparison: A performance comparison table will be designed based on the requirements provided by Enverus. The developed models will be evaluated and compared using relevant evaluation metrics. These metrics may include RMSE, R-squared, MAE, or other appropriate measures. The performance analysis will provide insights into the impact of different data sources and modeling techniques on the prediction of solar energy generation performance.

Through this comprehensive experimentation design, we aim to gain valuable insights into the impact of geographical locations and modeling techniques on the accuracy and reliability of solar energy generation prediction. By systematically analyzing and comparing different design options, we contribute to the advancement of accurate and reliable methodologies in the field of solar energy prediction. Ultimately, this knowledge will support informed decision-making processes and drive the adoption of sustainable energy solutions in the renewable energy sector.

# results aND VISUALIZATIONS

## Linear Regression

The initial model used in this study was a simple linear regression model that utilized only the scaled Global Horizontal Irradiance (GHI) values. The purpose of this model was to serve as a baseline for evaluating the performance of more complex models. The dataset was randomly divided into an 80% training set and a 20% testing set.

Upon evaluation, the simple linear regression model achieved an overall Root Mean Squared Error (RMSE) of 1744 MW and an R-squared (R2) value of 85.2% on the testing set. However, it is important to note that these results can be misleading due to the presence of approximately half of the records representing night-time hours with zero power generation. This imbalance in the dataset can significantly impact the model's performance.

To address this issue, the night-time hours (between 22:00 p.m. to 5:00 a.m. inclusive) were excluded from the dataset. After removing these zero production hours, the simple linear regression model was re-fitted. The revised model achieved an overall RMSE of 1997 MW and an R2 value of 80.2% on the testing set. This confirms our initial suspicion that including zero production hours can lead to optimistic and misleading results.

A scatter plot comparing the predicted values to the actual power generation on the testing set (Figure-2) indicates that the simple linear regression model is not suitable for accurately forecasting power generation. Additionally, the error distribution of the testing set does not follow a normal distribution, further supporting the observation that the simple linear regression model does not adequately fit the data (Figure-3 & 4).

Figures 3 and 4 provide a comparison between the actual and modeled solar power generation for a winter and a summer week in 2021 and 2022, respectively. These visualizations highlight the disparity between the actual and predicted values, demonstrating the limitations of the simple linear regression model in capturing the complex dynamics of solar power generation.

These findings underscore the need for more advanced and sophisticated modeling techniques to accurately forecast solar power generation. The limitations of the simple linear regression model serve as a motivation to explore alternative modeling approaches, such as machine learning algorithms and time-series models, which can better capture the non-linear relationships and temporal dependencies inherent in solar power generation data.

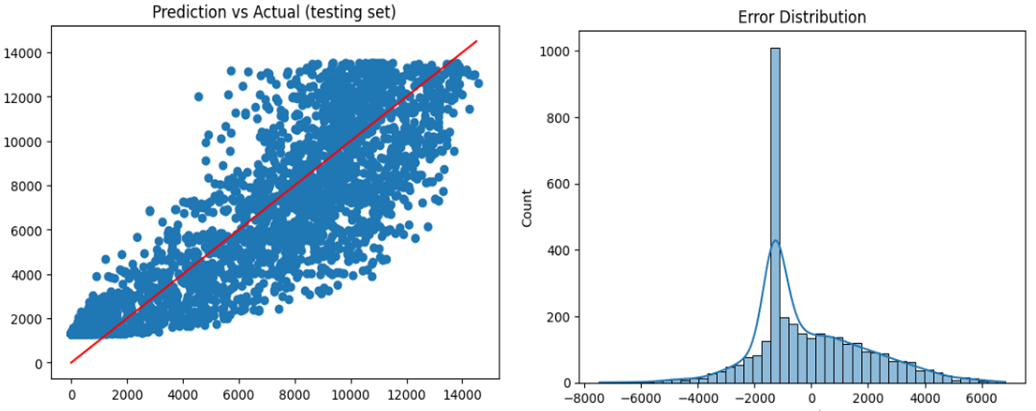


Figure-2. Prediction vs Actual; & error distribution for simple linear regression

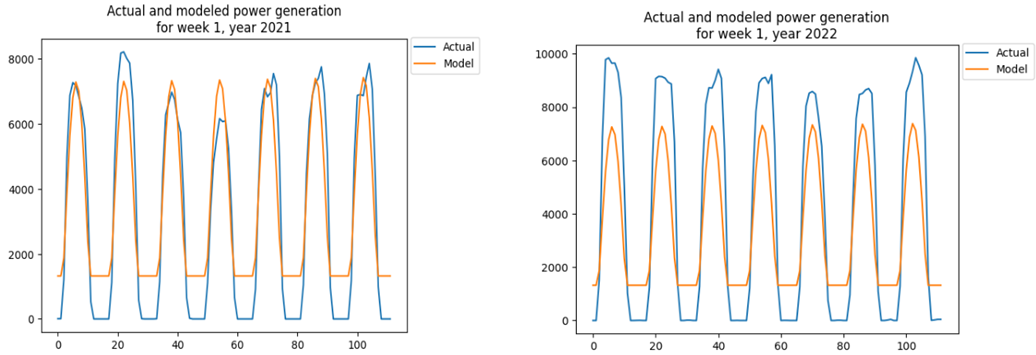


Figure-3. Actual and model power generation in a winter week in 2021 & 2022.

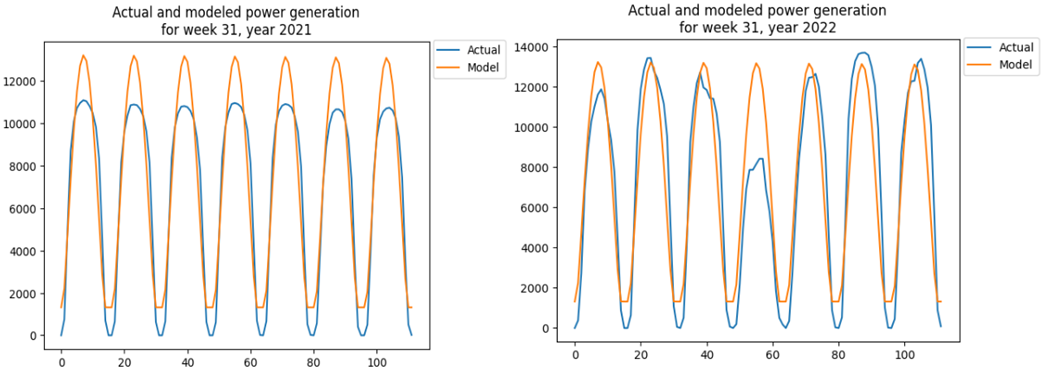


Figure-4. Actual and model power generation in a summer week in 2021 & 2022.

In this figure, we present a comparison between the actual and modeled solar power generation for a summer week in both 2021 and 2022. The x-axis represents the time period, while the y-axis represents the power generation in megawatts (MW). The blue line represents the actual power generation values, while the orange line represents the modeled power generation values.

The multiple linear regression (MLR) model, utilizing all numeric features except for the excluded night time hours, achieved an RMSE (Root Mean Squared Error) of 1492 MW and an R-squared value of 89.2% on the testing set. After feature selection, the final MLR model utilizing only GHI (Global Horizontal Irradiance) and DNI (Direct Normal Irradiance) achieved an RMSE of 1606 MW and an R-squared value of 87.2%.

The scatter plot in Figure-5 demonstrates the disparity between the predicted values and the actual power generation. The spread of the data points indicates that the MLR model may not be suitable for accurately forecasting power generation. Furthermore, the error distribution of the testing set does not appear to be normally distributed, confirming the earlier observation that the MLR model does not adequately fit the data.

Figure-6 and Figure-7 provide a visual comparison between the actual and modeled solar power generation for a winter and a summer week in 2021 and 2022. These graphs further highlight the differences between the actual and modeled values, emphasizing the limitations of the MLR model in accurately capturing the underlying patterns and dynamics of power generation.

Given the inadequate performance of the MLR model, it is necessary to explore alternative modeling approaches, such as random forest models, to improve the predictive accuracy and capture the non-linear relationships present in the data.

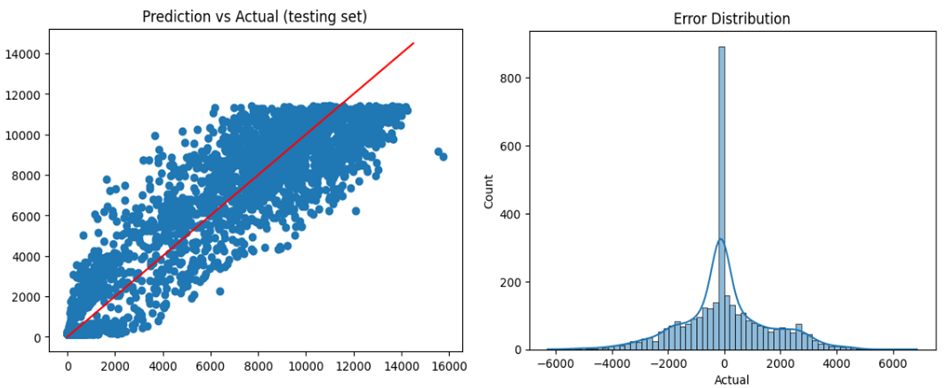


Figure-5. Prediction vs Actual; & error distribution for MLR model.

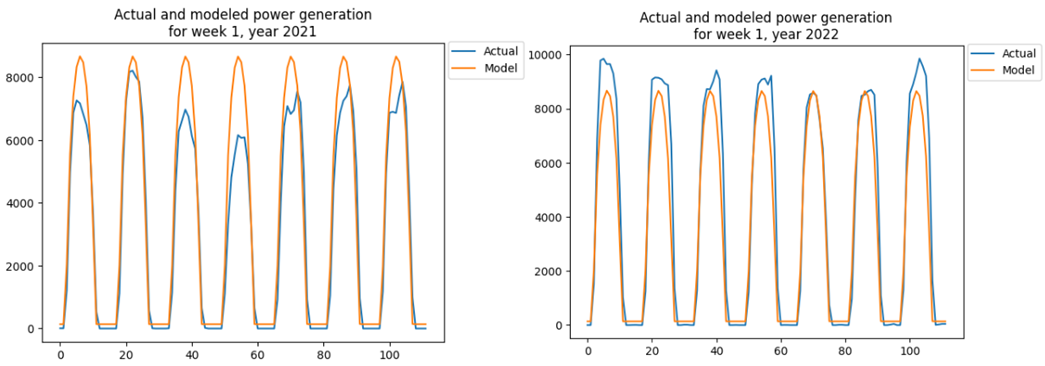


Figure-6. Actual and model power generation in a winter week in 2021 & 2022.

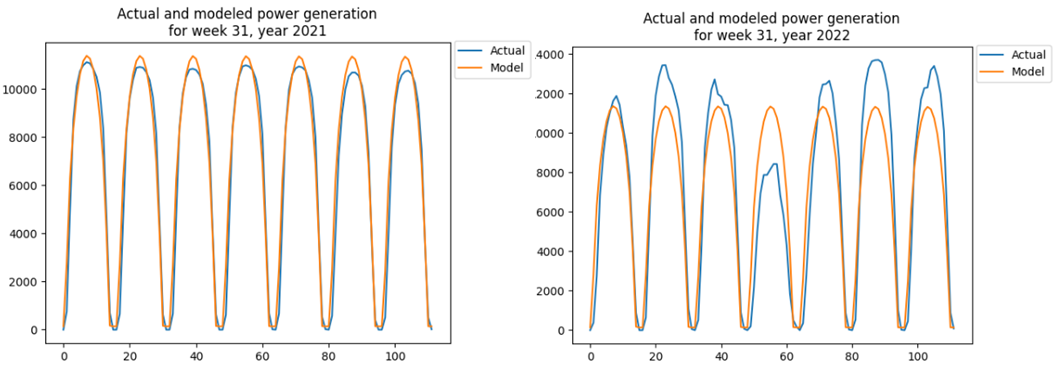


Figure-7. Actual and model power generation in a summer week in 2021 & 2022.

## Random Forest

Random Forest is a powerful and effective modeling technique for predicting solar farm performance due to its ability to handle complex relationships and capture nonlinearities in the data. It is an ensemble learning method that combines multiple decision trees, making it robust against overfitting and capable of handling high-dimensional datasets. Random Forest also provides feature importance measures, allowing us to identify the most influential factors affecting solar power generation. Its ability to handle a wide range of input variables and produce reliable predictions makes it a suitable choice for accurately forecasting solar farm performance.

The random forest model was developed using the scaled numeric features of the entire dataset, with categorical features excluded as their information is already captured in some of the numeric features. The initial model was built with 100 trees and the scaled numeric features. Subsequently, multiple model fit iterations were conducted by excluding one feature at a time to determine the significant features, assessed through R-squared value and RMSE.

The assessment identified GHI, DNI, cloud cover, humidity, pressure, and the week number as significant features for the random forest model. To optimize the model's performance, the number of trees was fine-tuned based on RMSE and R-squared value, as depicted in Figure-8. The optimal number of trees was determined to be 20, resulting in an overall RMSE of 1269 MW and an R-squared value of 92.2% on the testing set.

A scatter plot in Figure-9 displays the predicted values versus the actual power generation on the testing set, indicating that the random forest model is a suitable model for forecasting power generation. Furthermore, the error distribution of the testing set appears approximately normally distributed, confirming the earlier observation that the random forest model fits the data reasonably well.

Figure-10 and Figure-11 illustrate the comparison between actual and modeled solar power generation for a winter and a summer week in 2021 and 2022, demonstrating the satisfactory performance of the random forest model in capturing the underlying patterns and dynamics of power generation.

The superior performance of the random forest model suggests that it is a promising approach for accurately predicting solar farm performance.

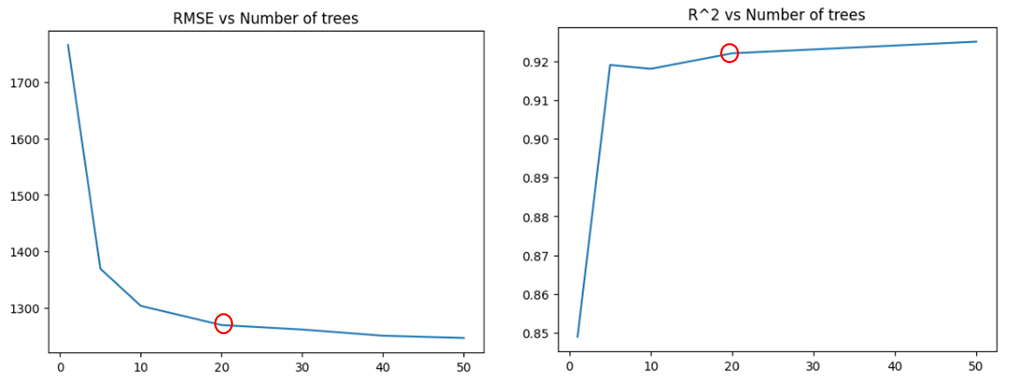


Figure-8. RMSE & R2 vs Number of Trees.

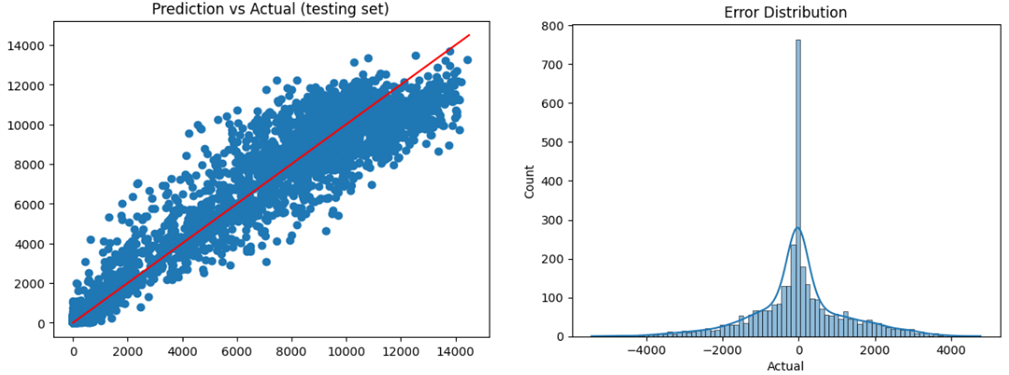


Figure-9. Prediction vs Actual; & error distribution for random forest model.

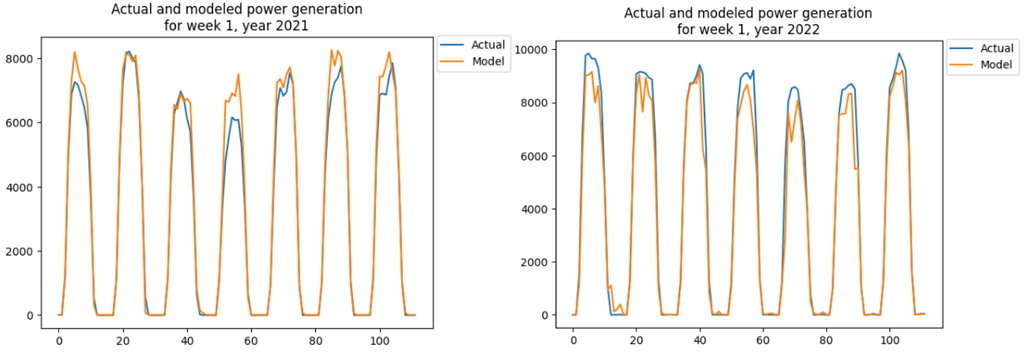


Figure-10. Actual and model power generation in a winter week in 2021 & 2022.

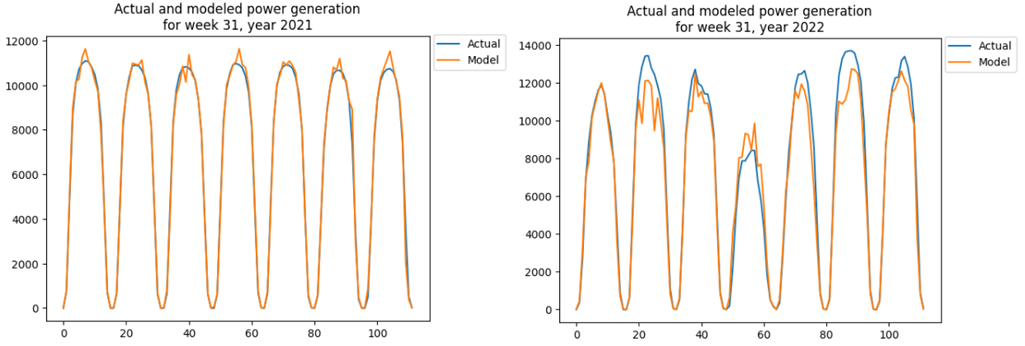


Figure-11. Actual and model power generation in a summer week in 2021 & 2022.

## Time Series Modeling

Time series models, such as autoregressive models, are used to capture the temporal patterns and dependencies in the data. They are particularly suited for forecasting tasks when the data has a clear sequential structure, as is the case with solar power generation. Time series models can capture seasonality, trends, and autocorrelation present in the data, allowing for accurate predictions of future values. However, in the specific context of this project, it was observed that the autoregressive model performed poorly in forecasting solar power generation. Additionally, time series models lack explainability, which is crucial for interpreting and understanding the factors influencing the forecasted values. As a result, it is concluded that autoregressive or time series models may not be suitable for the power generation forecast problem in this study.

Time series models, including autoregressive models, were explored to analyze the patterns and dependencies in the actual solar power generation data. The auto\_arima function was used to determine the optimal parameters for the ARIMA model. However, when using the suggested model order, the forecasts for one week ahead were found to be poor in performance. Further investigation revealed that only the autoregressive component of the ARIMA model was significant, indicating that the model could be simplified to an autoregressive model. The optimal number of lags for the autoregressive model was determined based on the partial autocorrelation analysis. To maintain an 80-20 train-test split, the autoregressive model was trained on 840 hours of data, estimating the parameters associated with each of the 74 lags, in order to forecast the future 168 hours.

However, the autoregressive model achieved an overall RMSE of 2015 MW and an R2 value of 76.9%, which is similar to the reference simple regression model. Additionally, time series models lack explainability, which is important for interpreting and understanding the forecasted values. As a result, it is concluded that autoregressive models, or time series models in general, may not be suitable for accurately forecasting power generation in this study. The forecast performance of the autoregressive model in two different weeks is depicted in Figure-14, highlighting the weak performance of the model.

A picture containing line, diagram, text, plot

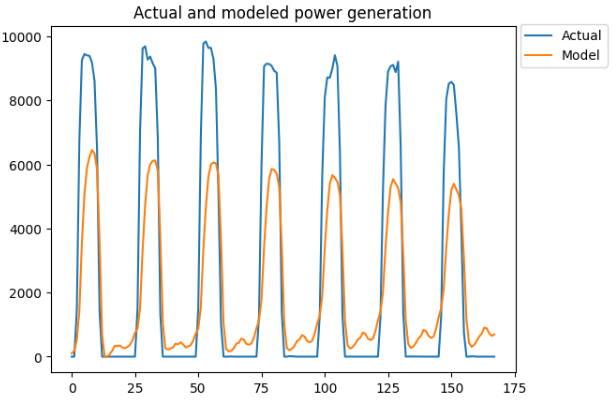
Description automatically generated

A picture containing text, line, screenshot, plot

Description automatically generated

Figure-11. Actual and modeled power generation. Figure-12. Autoregressive partial autocorrelation.

A picture containing text, screenshot, plot, line

Description automatically generated

A red line with blue dots

Description automatically generated with low confidence

A red line with blue dots

Description automatically generated with low confidence

Figure-14. Autoregressive model performance in 2 different weeks.

# conclusion and discussion

A comparison of the real solar power production and the modeled solar power generation for a week during the summer of 2021 and 2022 is shown in the image below. The y-axis depicts the amount of electricity produced in megawatts (MW), while the x-axis indicates the time period. The values of the actual power production are shown by the blue line, while the values of the power generation predicted by the model are represented by the orange line.

The difference between the predicted and observed levels of electricity production is shown graphically here. The multiple linear regression model that was employed for predicting the power generation does not adequately represent the underlying patterns and dynamics since there are notable discrepancies between the two lines. This indicates that the model should not be utilized. The multiple linear regression model may not be capable of providing an accurate prediction of power production due to the differences that exist between the real values and the predicted values.

These findings put an even greater emphasis on the need to investigate alternative modeling strategies that are capable of better capturing the intricacies of solar power production. In the parts that follow, we will talk about the investigation of random forest models as an alternate method to enhance the prediction performance and capture the non-linear correlations present in the data.

# references

1. Aksoy, N., & Genc, I. (2023). Predictive models development using gradient boosting based methods for solar power plants. Journal of Computational Science, 67(101958), 1-10. doi:https://doi.org/10.1016/j.jocs.2023.101958
2. California ISO. (2023). California ISO Open Access Same-time Information System. Retrieved from California ISO: <http://oasis.caiso.com/mrioasis/logon.do?reason=application.baseAction.noSession>
3. Chikate, B. V., & Sadawarte, Y. (2015). The factors affecting the performance of solar cell. International journal of computer applications, 1(1), 1-5. doi:0975 – 8887
4. Enverus. (2023). Intelligent Connections. Retrieved from Enverus: <https://www.enverus.com>
5. Hobbs, B. F., Zhang, J., Hamann, H. F., Siebenschuh, C., Zhang, R., Li, B., . . . Zhang, S. (2022). ISO, Using probabilistic solar power forecasts to inform flexible ramp product procurement for the California. Solar Energy Advances, 2(100024), 1-11. doi:https://doi.org/10.1016/j.seja.2022.100024
6. Miao, S., Ning, G., Gu, Y., Yan, J., & Ma, B. (2018). Markov Chain model for solar farm generation and its application to generation performance evaluation. Journal of Cleaner Production, 186(1), 905-917. doi:https://doi.org/10.1016/j.jclepro.2018.03.173
7. Singh, A. K., & Singh, R. R. (2021). An overview of factors influencing solar power efficiency and strategies for enhancing. Innovations in Power and Advanced Computing Technologies (i-PACT), 1(1), 1-6. doi:10.1109/i-PACT52855.2021.9696845

# Appendices

Appendix 1

Enverus File

A picture containing screenshot, plot

Description automatically generated

A picture containing screenshot, diagram, design

Description automatically generated

Records main file : 26,193 x 3 columns / 463 KB

Appendix 2

Merged File used for the models



Records merged file : 26,193 x 22 columns / 5.38MB