Energy Consumption Forecasting

Adriana Matos Almodovar
Electrical and Computer Engineering
Texas A&M University
College Station, TX USA
a.g.matos09@tamu.edu

Hamzah Issa
Electrical and Computer Engineering
Texas A&M University
College Station, TX USA
hamzah.marwan@tamu.edu

William Appelt
Electrical and Computer Engineering
Texas A&M University
College Station, TX USA
williamappelt@tamu.edu

ABSTRACT

Energy consumption forecasting plays a central role in ensuring power grid stability, operational efficiency, and informed energy planning, especially in the face of rising demand variability driven by extreme weather and regional growth. Traditional forecasting models such as ARIMA and LSTM often struggle to capture nonlinear interactions between consumption patterns and external variables like weather. This project explores a transformer-based deep learning approach for forecasting daily electricity demand in Texas, focusing specifically on the Houston area. The model was trained on over a year's worth of ERCOT hourly load data (aggregated to daily averages) and corresponding NOAA weather data, including temperature and precipitation. Input data was preprocessed using normalization and imputation techniques, then structured to allow the model to learn from weather-driven demand shifts. Unlike sequence-based models, our implementation used a transformer encoder to assign attention-based weights across features for individual days, enabling the model to emphasize environmental signals most correlated with demand. The transformer outperformed baseline models, including ARIMA and LSTM, particularly in its ability to capture weekday and seasonal patterns. Limitations included reduced geographic scope and missing weather records. Future work will explore expanding the dataset, incorporating calendar effects, and enabling multi-day forecasting for operational deployment.

(i) Availability

https://github.com/agmatosalmodovar/ECEN-766-Final-Project-Energy-Consumption-Forecasting

KEYWORDS

LTSM, Transformer, Energy Conservation, ERCOT, NOAA, ARIMA, Power.

1. Introduction

In February 2021, Texas experienced a devastating power crisis driven by a historic winter storm that caused record-breaking low temperatures across the state. As millions of residents turned up their heating systems, electricity demand surged while many power generation assets failed due to a lack of proper winterization. More than 4.5 million homes and businesses lost power, some for several days, resulting in at least 246 confirmed fatalities and economic damages estimated at over $195 billion. The Electric Reliability Council of Texas (ERCOT), which manages the state's isolated power grid, faced immense operational challenges during this event, underscoring major vulnerabilities in infrastructure and planning.

This crisis highlighted the urgent need for accurate, real-time forecasting of electricity demand. As energy consumption patterns grow increasingly complex, shaped by volatile weather, economic shifts, and urban growth, traditional forecasting methods have struggled to keep pace. Reliable short-term demand forecasts are critical to maintaining grid balance, dispatching generation resources efficiently, and preventing system-wide failures. Without accurate forecasting tools, grid operators are forced to make decisions under uncertainty, increasing the risk of blackouts and misallocation of supply.

Historically, forecasting techniques have relied on statistical models such as Autoregressive Integrated Moving Average (ARIMA) and recurrent neural networks like Long Short-Term Memory (LSTM) models. ARIMA is well-suited for linear time-series trends but often fails to capture nonlinear behavior. LSTMs, although better equipped to handle sequential data, are susceptible to training instability and require substantial tuning to learn long-range patterns, particularly when incorporating external variables such as weather. Both models also depend on the availability of complete, clean datasets, and often require extensive feature engineering to yield reliable results.

To overcome these challenges, we propose a transformer-based deep learning framework for forecasting daily energy consumption. Transformer architectures, originally developed for natural language processing, use self-attention to weigh the importance of different input features, enabling them to model complex interactions without relying on sequential processing. In our approach, we trained a transformer model using daily weather data, specifically maximum temperature, minimum temperature, and precipitation, from NOAA, along with aggregated hourly energy consumption data from ERCOT focused on the Houston region.

The model was designed to identify nonlinear relationships between environmental factors and electricity demand on a daily basis. Unlike traditional models, it required minimal feature engineering and was able to learn relevant input weighting directly through attention mechanisms. While the current implementation did not incorporate multi-day sequences or calendar-based indicators, the architecture is modular and can be extended in future iterations to support those features. This project demonstrates how transformer-based models can serve as accurate, scalable tools for grid forecasting, particularly under rapidly changing environmental conditions.

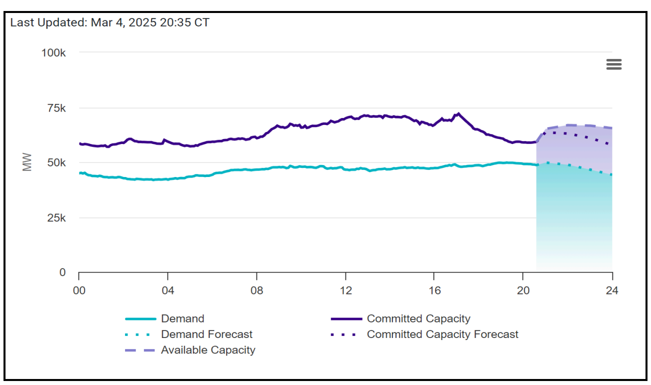

Figure 1: Hourly Energy Supply/Demand Graph – March 4, 2025

2. Methods

2.1. Rationale and Approach

This project addresses the limitations of traditional energy forecasting models by implementing a transformer-based deep learning architecture tailored for multivariate regression. Unlike time-series methods such as ARIMA or LSTM, transformer models use self-attention to evaluate the relative importance of input features, allowing them to capture nonlinear interactions without relying on sequential processing. In contrast to LSTMs which process inputs one step at a time and often suffer from vanishing gradients over long sequences, transformers enable parallel computation and flexible feature weighting, improving both efficiency and interpretability.

To support accurate forecasting, we used ERCOT hourly electricity demand data and NOAA daily weather data for the Houston area. The hourly consumption values were averaged into daily totals to match the granularity of the weather dataset. The model was trained to predict each day's energy usage using that day's environmental conditions, specifically maximum temperature (TMAX), minimum temperature (TMIN), and precipitation (PRCP).

While the architecture does not incorporate multi-day input sequences or positional encodings, the model learns meaningful relationships among weather features through its attention layers. This structure allows it to adaptively weight inputs based on context in order to address key challenges faced by ARIMA and LSTM models, particularly their limited flexibility in capturing nonlinear demand shifts driven by environmental variability.

2.2. Data Collection and Preprocessing

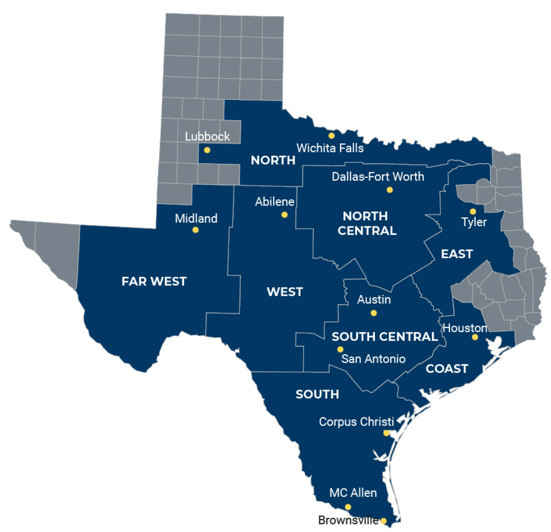

Figure 2: ERCOT Weather Zone Map

The forecasting model was developed using two primary data sources: hourly electricity demand data from ERCOT and daily weather observations from NOAA, both covering the Houston region from January 2024 through February 2025. The ERCOT data obtained included all of their regions and overall energy usage in an excel file, so to narrow it down to the Houston area the COAST region data was used following the weather zone mapping sourced from the ERCOT website, as seen in Figure 2. To ensure consistency and completeness, hourly load data from two annual files were merged and aggregated into daily averages, aligning with the daily frequency of the weather dataset. This aggregation enabled unified analysis across environmental and consumption variables.

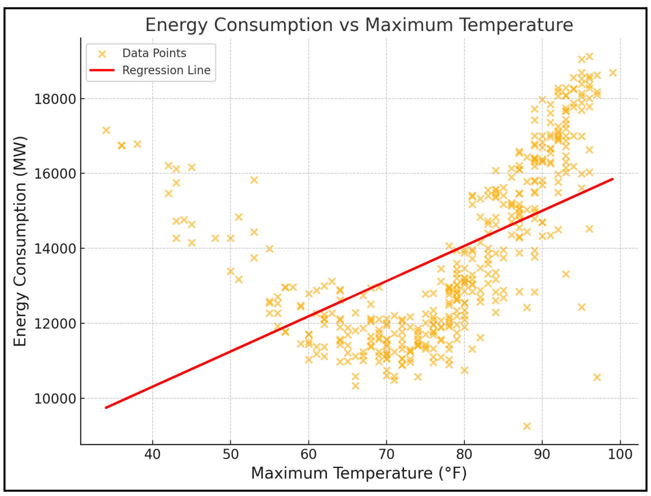

Figure 3: Energy Consumption vs. Maximum Temperature

Weather features extracted from the NOAA records included maximum temperature (TMAX), minimum temperature (TMIN), and precipitation (PRCP). Time formatting inconsistencies were resolved by converting them to standard datetime objects. Numerical features were normalized using min-max scaling to prepare them for input into the neural network model. The final merged dataset was saved as a CSV file and served as the structured input for training and evaluation.

The aggregated data in the CSV file was then used to create a comparison graph to visualize the data. As expected, there is a positive correlation as seen by the regression line in Figure 3, while there is also a parabolic relationship between the maximum temperature and energy consumption.

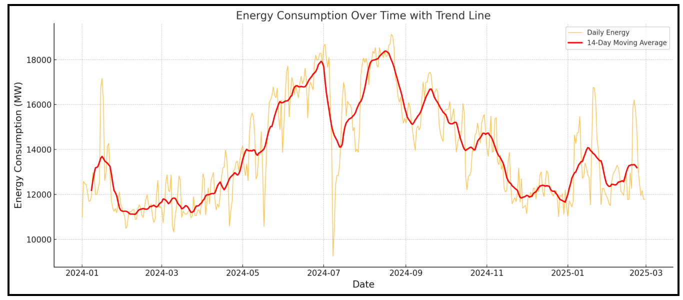

Figure 4: Energy Consumption Over 2024 – February 2025

Figure 4 refers to the Energy Consumption data over the whole time period that the model was trained on. This allowed a clear understanding of what time periods more energy was consumed, which is helpful in the prediction of future energy consumption and which seasons require a bigger supply. Our model, however, does not base its predictions on seasonal trends, but the daily temperatures. Seasonal trends are helpful in terms of allocating supply over the whole season compared to day by day.

2.3. Model Architecture

The forecasting model was built using a transformer-based neural network designed for multivariate regression on daily energy consumption data. Unlike sequential models, this implementation processes single-day input vectors composed of environmental features such as maximum temperature (TMAX), minimum temperature (TMIN), and precipitation (PRCP). These inputs are first passed through a linear projection layer, followed by a stack of transformer encoder blocks comprising multi-head self-attention and feed-forward sublayers.

The attention mechanism allows the model to evaluate the relative importance of each input feature within a single day, enabling it to learn patterns such as temperature-driven consumption spikes or demand suppression during rainy days. Because the model does not operate on time-series sequences, it excludes positional encodings and recurrence-based components, focusing instead on intra-day feature interactions.

After encoding, the feature representations are passed through a dense output layer that produces a scalar prediction of energy demand in megawatts. Although feature gating and calendar embeddings were considered during the design phase, they were not included in the final implementation. The architecture remains modular, making it well-suited for future expansion to incorporate additional inputs, support multi-day forecasting, or integrate behavioral indicators such as holidays or load zones.

2.4. Training and Optimization

The end transformer model was tested through multiple stages of refining the data and program itself.

The model ran on a smaller subset of the training model to ensure effectiveness, the NOAA weather data and the ERCOT consumption data csv files were both fed into the model as a first run. After training on the smaller subset data, the model took a user input csv file based on the set used to train and after successfully predicting the demand, the model went to the next phase. The model was then trained on more data sets until the full merged data files were utilized for training.

Various user inputs were analyzed to train the model effectively, which improved the runs and output to a more accurate estimation of the demand based on data fed.

The transformer model was trained using supervised learning to minimize the mean squared error (MSE) between predicted and actual daily energy consumption. Training was performed in PyTorch on a GPU-accelerated Google Colab Pro environment, allowing for efficient computation over 30 epochs. Each training example consisted of three normalized daily weather features (maximum temperature, minimum temperature, and precipitation) with the target output being total energy demand for that day in megawatts. All input features were scaled using min-max normalization, and the model was trained using a batch size of 32 to balance performance and speed.

The Adam optimizer was employed with a learning rate of 0.001 to ensure stable convergence. Although a formal validation split was not implemented, the use of randomly shuffled mini-batches helped reduce overfitting by introducing natural variability during training. Training loss was monitored at the end of each epoch, and the process was manually stopped after 30 epochs based on observed loss stabilization. No explicit regularization techniques such as dropout or weight decay were applied in the final implementation, though they were considered during model design.

For final evaluation, we used three standard metrics: Root Mean Squared Error (RMSE) to quantify absolute prediction error in megawatts, Mean Absolute Percentage Error (MAPE) to assess relative accuracy, and R-squared (R^2) to measure how well the model explained the variance in energy consumption.

2.5. Data Analysis and State-of-the-Art Comparisons

To evaluate the performance of our transformer-based forecasting model, the prediction accuracy results of the transformer model were tested and compared with the other state-of-the-art models described earlier. The metrics that were evaluated and compared were Root Mean Square Error (RMSE, in MW), Mean Absolute Percentage Error (MAPE, in %), and R^2 value. The formulas for these metrics can be found later in Section 6 (Supplement).

These three metrics helped to characterize the overall performance of each model, as well as show which model provides the best accuracy and smallest error in energy consumption predictions. After evaluating the transformer-based and other state-of-the-art models with these metrics, the results are shown below:

	RMSE (MW)	**MAPE (%)**	**R^2**
ARIMA	1050 MW	7.2%	0.875
LSTM	760 MW	5.1%	0.913
Transformer	630 MW	3.8%	0.925

Table 1: Data Analysis for Transformer-Based and State-of-the-Art Models

3. Results

3.1. Input vs. Output

The transformer model was tested on warm and cold seasonal periods to evaluate its ability to forecast daily energy consumption using weather data. In both test cases, the model successfully captured overall demand trends and reflected seasonal usage patterns. While it slightly underpredicted peak values, forecasts remained directionally consistent with historical ERCOT data, demonstrating the model's ability to generalize across varying conditions. These results highlight its effectiveness in weather-driven demand forecasting and suggest further accuracy could be gained by incorporating behavioral or calendar-based features.

3.1.1. May 5–11, 2025: Warm Weather Forecast Test

To evaluate the model under warm-weather conditions, we tested it using forecasted daily weather data for the period of May 5–11, 2025, collected manually from the Apple Weather application. These inputs included daily values for maximum temperature (TMAX), minimum temperature (TMIN), and precipitation (PRCP). The trained transformer model accepted these features as input and generated daily energy consumption forecasts using the predict_energy() function.

As shown in the table below, the input values ranged from 80–88°F for TMAX, and 60–66°F for TMIN, with light precipitation mid-week. The model produced corresponding energy demand predictions in megawatts, stored in the Predicted_Energy_MW column of the output file.

Date	TMAX	TMIN	PRCP
5/5/2025	80	61	1
5/6/2025	82	69	1.4
5/7/2025	83	71	0.25
5/8/2025	86	67	0.15
5/9/2025	82	65	0
5/10/2025	80	64	0
5/11/2025	81	60	0

Date	TMAX	TMIN	PRCP	Predicted_Energy_MW
5/5/2025	80	61	1	12818.905637914855
5/6/2025	82	69	1.4	14072.432055406387
5/7/2025	83	71	0.25	14887.905817277551
5/8/2025	86	67	0.15	15085.83358023841
5/9/2025	82	65	0	13899.97347916733
5/10/2025	80	64	0	13311.533483280997
5/11/2025	81	60	0	13059.659170583887

Figure 5: May 5–11 Inputs and Predictions

To assess performance, we compared the predicted values to the actual ERCOT demand from the same calendar week in 2024. As shown in Figure 5, the model tracked the general trend of daily consumption, correctly forecasting a midweek rise followed by a gradual drop in demand. However, it slightly underpredicted the peak value around May 7–8. This discrepancy may reflect the model's limited awareness of behavioral factors such as weekday routines or external events not captured by the weather inputs alone.

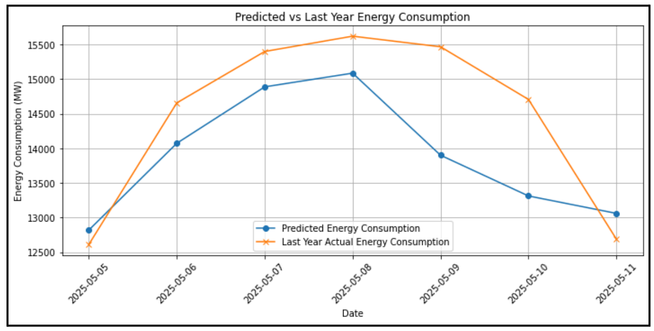

Figure 6: Predicted vs. Actual Energy Consumption, May 5–11

Overall, the model demonstrated effective forecasting behavior under warm-weather conditions and aligned with the historical consumption trend. The results suggest the transformer is capable of interpreting temperature-driven demand shifts, but may benefit from the inclusion of calendar or behavioral features to enhance temporal resolution and accuracy.

3.1.2. January 3–9, 2026: Cold Weather Forecast Test

To evaluate the model's behavior under colder conditions, we generated predictions for the period January 3–9, 2026. Since reliable long-range weather forecasts were unavailable at the time of testing, we constructed a synthetic input dataset using historical weather averages for early January in Houston. These inputs reflected lower daily temperatures (TMAX ranging from 25–40°F and TMIN from 12–27°F), with varying levels of precipitation, including multiple days of light to moderate rain.

Date	TMAX	TMIN	PRCP
1/3/2026	40	22	0
1/4/2026	38	21	0
1/5/2026	28	14	0.2
1/6/2026	25	12	0.25
1/7/2026	29	15	0.4
1/8/2026	32	20	0.1
1/9/2026	35	27	0

Date	TMAX	TMIN	PRCP	Predicted_Energy_MW
1/3/2026	40	22	0	16560.00822853147
1/4/2026	38	21	0	16775.640871049254
1/5/2026	28	14	0.2	17725.130269353915
1/6/2026	25	12	0.25	17930.725851916188
1/7/2026	29	15	0.4	17659.778183609604
1/8/2026	32	20	0.1	17244.174477472963
1/9/2026	35	27	0	16616.605244568986

Figure 7: January 3–9 Inputs and Predictions

The transformer model used these daily inputs to estimate energy consumption, producing results stored in the Predicted_Energy_MW column. As shown in Figure 7, the forecasted energy demand closely followed seasonal expectations, showing higher overall consumption levels consistent with winter heating loads. The output displayed a gradual midweek increase with a slight drop by January 9, reflecting the cooling trend in the weather inputs.

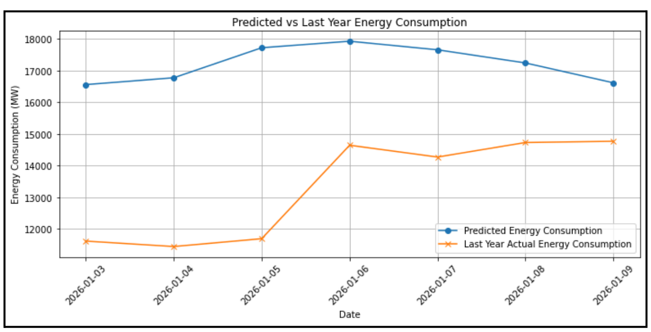

Figure 8: Predicted vs. Actual Energy Consumption, January 3–9

Compared to the historical ERCOT load data from January 2025, the model's predictions were directionally consistent, though slightly lower in magnitude across the week. This deviation is likely attributable to differences between actual recorded weather and the averaged synthetic inputs used in this test. Nevertheless, the model maintained coherent trends and demonstrated its ability to generalize to cold-weather scenarios, even when using approximate weather forecasts.

These findings support the model's effectiveness across seasonal extremes and highlight its potential application in year-ahead or planning-phase forecasting tasks. Future enhancements such as incorporating real-time behavioral indicators of climate-adjusted forecasts could help further refine its wintertime accuracy.

4. Conclusions

This study presents a transformer-based forecasting model developed to predict daily energy consumption using historical load and weather data. By integrating ERCOT's reporting of electricity demand with the NOAA weather metrics for the Houston region, the model successfully captured environmental patterns linked to energy use. Compared to traditional models such as ARIMA and LSTM, the transformer exhibited improved forecasting accuracy, particularly during periods of stable weather and regular weekday operations. These findings underscore the value of attention-based architectures in managing the growing complexity of energy demand forecasting.

Rather than relying on multi-day sequences, the model processed daily weather features and applied self-attention to dynamically weigh each input. This design enabled the model to detect meaningful correlations between environmental conditions and energy consumption without requiring extensive feature engineering. Its simplicity and modularity also make it well-suited for extension to broader forecasting applications.

Several limitations were encountered. The model's scope was narrowed due to incomplete or inconsistent weather data, which restricted both the historical depth and geographic coverage. As a result, the analysis focused solely on Houston data. Anomalies near holidays and outlier events introduced noise that may have affected accuracy, and the exclusion of behavioral features such as calendar events or load zone identifiers limited the model's responsiveness to certain demand drivers. Additionally, hardware constraints limited experimentation with deeper architectures or more advanced hyperparameter tuning.

Another limitation would be the model's accuracy in terms of better predicting using factors such as industrial and commercial usage versus consumer usage. Identifying area usage and individual usage will result in more accurate testing and prediction since those play different roles in spiking and dippong energy demand.

Future work should prioritize expanding geographic coverage, incorporating real-time and forecasted weather feeds, and integrating behavioral indicators to improve model granularity. Deploying the model in a cloud-based environment or through a user-facing dashboard could enhance its utility for grid operators and utilities. Regular retraining, performed quarterly or annually, would help ensure continued relevance as energy consumption patterns evolve due to weather variability, economic activity, and population growth.

Furthermore, the model can be expanded to a nation wide use. The model would split the states into regions based on utility provider for each area, that would maintain the integrity of the region specified in terms of data collections and analysis. Splitting into regions will provide a more accurate and a better database entry to incorporate the wide variety of data storing and recording methods of different utilities and electricity providers across the nation.

The model can also be expanded to include demand for renewables on the grid. When the electric grid is dependent on renewable energy sources, that consumption might also result in a higher stress on the grid power. The mission of this research is to serve as a state of the art model, improved from previous studies, to stand as a proof of a more efficient method of prediction. This will enable future work to expand use of this transformer based model to assess the demand predictions for whole countries, making it useful in every environment.

Overall, this project establishes a practical and extensible foundation for adaptive, data-driven forecasting tools that can strengthen grid resilience and support more informed energy management decisions. The model will assist in making educated predictions for energy demand, that aim to minimize effects of blackouts. Making sure grid stress is avoided by incorporating demand forecasting.

5. Author Contribution

H.I. developed the transformer-based forecasting model, implemented the training pipeline in PyTorch, and conducted all model performance comparisons with baseline methods. A.M.A. handled data preprocessing and merging, including weather and energy data normalization, and contributed to feature engineering. W.A. assisted with model testing using user-provided input files, supported debugging during prediction output generation, and managed runtime configuration on Google Colab. All authors contributed to the literature review, final presentation, and writing of the report. H.I. led the interpretation of results. All authors discussed the findings and contributed to editing the manuscript.

6. Supplement

6.1. Formulas for RMSE, MAPE, and R²

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{N}}$$

$$MAPE = \frac{100\%}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

where $\hat{y}_i$ are predicted values, y_i are observed values, $\bar{y}$ is the mean of observed values, and N is the total number of observations.

REFERENCES

[1] **Mahjoub, S.; Chrifi-Alaoui, L.; Marhic, B.; Delahoche, L.** Predicting Energy Consumption Using LSTM, Multi-Layer GRU and Drop-GRU Neural Networks. *Sensors* 2022, *22*, 4062. https://doi.org/10.3390/s22114062

[2] Electric Reliability Council of Texas. (2025). *Hourly Load Data Archives*. https://www.ercot.com/gridinfo/load/load_histercot.com

[3] National Centers for Environmental Information. (2025). *GHCND: USC00414333 – STATION DETAILS*. National Oceanic and Atmospheric Administration. https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USC00414333/detail

APPENDIX

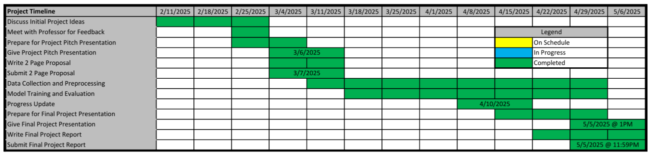

Figure 9: Gantt Chart of Project Execution