Enhancing Utility Weather Station Network Measurements by Identifying Discrepancies With North American Mesoscale Data and Polygon Optimization

Kenneth Hidayat
khidayat@ucsd.edu

Steven Sahar
ssahar@ucsd.edu

Phi Nguyen
pnguyen@sdge.com

Kasra Mohammadi
KMohamma@sdge.com

Jacob Wigal
JWigal@sdgecontractor.com

Abstract

To ensure that Public Safety Power Shutoffs (PSPS) are effectively implemented when necessary, it is crucial to obtain accurate wind speed readings. Weather station data serves as the ground truth, while NAM forecasts provide an additional reference to evaluate and confirm those readings. However, discrepancies between these sources can create uncertainty in critical decision-making, highlighting the need for a systematic approach to assess and address these inconsistencies. Previous research has used forecasting models, such as the Weather Research and Forecasting (WRF) model, to identify inconsistencies between model predictions and weather station measurements. However, these studies have not focused on pinpointing specific locations with significant discrepancies. This paper aims to conduct a geospatial analysis and apply loss functions to identify North American Mesoscale (NAM) points with large wind speed errors compared to wind speed measurements from SDG&E weather stations. Since error measurements can only be accurately determined for points within VRI polygons, we utilized the LightGBM model to predict the error of NAM points located outside these polygons. Additionally, we seek to refine the VRI polygon boundaries by analyzing polygons with high NAM point discrepancies. Specifically, we will perform an in-depth evaluation of NAM points within these polygons and propose the closest alternative VRI polygons that reduces overall discrepancies. In general, our approaches enable the identification of locations where additional wind speed data should be collected and improve accuracy when executing wildfire mitigation strategies.

Repository: https://github.com/StevDoms/SDGE_NAM_Analysis

1 Introduction

The recent Palisades Wildfire and the San Diego Border Wildfires are prime examples of wildfires increasingly becoming a threat to communities, ecosystems, and infrastructure caused by mixed factors ranging from climate change to human activities. Since 2007, SDG&E has installed infrastructures such as weather stations to strengthen precautionary and mitigatory measures against wildfires. Having precise weather data, particularly wind speed data, is crucial in evaluating wildfire risks and plays a significant role in the implementation of mitigation strategies, such as Public Safety Power Shutoffs (PSPS). The decision to conduct a PSPS is based on weather station readings, particularly wind speeds. Each weather station is associated with a Vegetation Risk Index (VRI) polygon, which represents an area where the station's readings are assumed to uniformly reflect local wind conditions. These polygons are designed by SDG&E meteorologists, who consider factors such as topography, historical weather patterns, and vegetation density to ensure that each polygon accurately captures wind behavior in its respective region.

To ensure that PSPS is effectively implemented when necessary, it is crucial to obtain more accurate wind speed readings. This would be especially true with larger VRI polygons that might face accuracy challenges in areas which are located further away from the weather station. Our project aims to address these challenges by identifying regions with inconsistent wind speed measurements, particularly where there is a high discrepancy between wind speed data from the North American Mesoscale (NAM) model and weather station measurements within VRI polygons. These areas of interest will provide valuable information for future evaluations regarding the collection of additional wind speed data. To account for NAM points located outside VRI polygons, we incorporate a LightGBM model to estimate their discrepancies, enabling a more comprehensive assessment of wind speed errors. We show the degree of these errors in the methods to allow a better understanding of the properties of these errors. Furthermore, our project focuses on refining VRI polygon boundaries by leveraging insights from NAM points within the polygons with high inconsistent wind speed measurements. We aim to optimize boundary assignments by associating NAM points having the top 20 highest discrepancies within a polygon with the nearest alternative VRI polygons to reduce wind speed errors. The new assignments of the polygons will serve as a basis for future boundary drawings.

Our initial Exploratory Data Analysis (EDA) of wind speeds indicates that NAM points tend to underestimate wind speeds compared to weather station measurements, with a correlation coefficient of 0.58, suggesting a moderate correlation. Upon overlaying NAM point discrepancies on a geospatial map, we identified several regions with high errors, particularly VRI polygon areas covering valleys around San Diego. These NAM points and VRI polygons in particular also follow the shapes of the valleys.

A previous study that applied a methodology similar to ours in relation to the NAM model is titled *"Examination of Errors in Near-Surface Temperature and Wind from WRF Numerical Simulations in Regions of Complex Terrain"* (Zhang, Pu, & Zhang, 2013). This paper evaluates the performance of the Weather Research and Forecasting (WRF) model, which conducts high-resolution local weather forecasting and utilizes the NAM model as one of

its input dataset. The study employs a methodology similar to ours, using surface weather station data as ground truth to compute the Mean Absolute Error (MAE) of the WRF model outputs. However, while this paper shares methodological similarities with our project, its primary focus is on analyzing trends in WRF model discrepancies rather than identifying the specific locations of these discrepancies. In contrast, our study explicitly evaluates spatial discrepancies between wind speed predictions from the NAM model and weather station measurements, aiming to pinpoint regions with significant forecast errors. Additionally, we extend this analysis by optimizing VRI polygon boundaries based on NAM point errors, introducing a novel approach to enhancing the spatial accuracy of wildfire risk assessments.

Our project utilizes data provided by SDG&E, which includes weather station locations, VRI polygons (electric asset mapping), and customized NAM forecasts created in partnership with UCSD SDSC which can be found here. We used the NAM forecast and wind speed data that are collected from weather stations owned by SDG&E. The NAM forecast is stored in NetCDF format, which supports multidimensional data structures. To ensure consistency, we filter the dataset for dates when PSPS was considered for implementation. These selected dates correspond to those available in our weather station wind speed dataset. The NAM wind speed data serve as a comparative dataset to identify discrepancies between existing weather station measurements and NAM model predictions. To ensure the accuracy of our geospatial analysis, all the data geometry must share the same coordinate reference system (CRS). Standardizing the CRS ensures precise alignment of weather station coordinates (Points), VRI polygons (Polygons), and NAM points (Points). Once the CRS is unified, we can systematically assess discrepancies between the weather station data and NAM model predictions, enabling us to compute wind speed errors and proceed with optimizing VRI polygon boundaries.

2 Methods

2.1 Obtaining the Data

To perform the data analysis, we must first prepare all required datasets. Our project utilizes weather station data, weather station wind speed data, VRI polygon data, conductor span data, NAM wind speed data, elevation data, San Diego County boundary and Southern Orange County boundary. Four of the latter datasets must be obtained from various external sources, which will be outlined in the following subsections.

2.1.1 NAM Wind Speed

The NAM wind speed data was compiled from here. The NAM model records wind speeds at an hourly interval. However, since the weather station wind speed data corresponds to a single measurement per day, we compute a simple daily average for each NAM wind speed point:

$$\text{Daily Average Wind Speed} = \frac{1}{N}\sum_{i=1}^{N} v_i$$

where v_i represents the wind speed recorded at hour i, and N is the total number of hourly recordings in a day.

The range of NAM data collected is from 2012 to 2024; however, we do not compute wind speed data for all the days within this range. Instead, we extract a subset of unique dates that match those in our weather station wind speed dataset—these dates correspond to days when PSPS was considered for implementation. In total, there are 179 unique dates.

The raw NAM data is stored in NetCDF (*.nc*) format, which is optimized for multidimensional data storage. Since our objective is to extract and analyze the average wind speed for the selected dates, we transform the dataset into a tidy format, where each row corresponds to a specific date and NAM point. To facilitate further analysis, we converted the compiled data from NetCDF format into a CSV file for easier processing and accessibility.

2.1.2 Weather Station & NAM Points Elevation

After processing the NAM data, we utilize the longitude and latitude of both the weather station and NAM points to obtain the elevation at those locations. This is achieved using the Open Elevation API, which returns the elevation for each specified point. The API call is structured as follows:

https://api.open-elevation.com/api/v1/lookup?locations={lat},{lon}

The elevation data obtained from this API will be used as one of the features for the LightGBM model.

2.1.3 County Boundary

One of the data preprocessing steps we performed to obtain relevant NAM points outside the VRI polygons was filtering the NAM data for points within SDG&E's area of operation, which covers San Diego County and the southern part of Orange County.

The San Diego County boundary data was obtained from here, while the Orange County boundary data was obtained from here. For the Orange County boundary, we specifically use the Congressional boundary as it matches SDG&E's operation boundary.

2.2 Data Processing & Wind Speed Error Calculation

To calculate the **wind speed error between each NAM points within the VRI polygon** and the weather stations, we will merge the weather station data, wind speed data, NAM data

and the VRI polygon data into a singular dataset before then calculating the Mean Absolute Error and Distance Weighted Absolute Error to calculate the degree of the discrepancies.

2.2.1 Data Merging and Filtering

The first step in merging the datasets involves combining the weather station data with the wind speed data by merging them on the *weatherstationcode*. This ensures that each weather station entry is paired with its corresponding wind speed measurement.

Next, we filter the NAM data to include only the points that fall within the VRI polygons. This is achieved through a spatial join using the geographic coordinates of the NAM points and the VRI polygon boundaries.

Following this, we perform another spatial join between the previously merged weather station dataset and the filtered NAM dataset. This step ensures that all data are consolidated into a single table, where every entry corresponds to a location within the VRI polygon.

One challenge in merging these datasets is the potential misalignment of dates between the NAM points and the weather station measurements. Since each dataset may have different timestamps, we apply a filtering step to retain only the entries where the NAM data and the weather station data have matching dates. This alignment is crucial to ensure the validity of our analysis and to accurately compute discrepancies between NAM wind speed predictions and actual weather station measurements.

2.2.2 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) quantifies the average absolute difference between the wind speeds of the NAM model and the wind speeds of the weather stations:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\text{NAM\_wind\_speed}_i - \text{station\_wind\_speed}_i\right|$$

where:

- NAM_wind_speed_i represents the NAM model's wind speed at a specific NAM point and date i.
- $\text{station\_wind\_speed}_i$ represents the observed wind speed at the corresponding weather station for date i.

This MAE error function is used to evaluate the discrepancies of each NAM point of those **NAM points within the VRI polygons** to later **predict the MAE error of NAM points outside the VRI polygons**.

2.2.3 Distance-Weighted Absolute Error (DWAE)

The Distance-Weighted Absolute Error (DWAE) calculates the product of the absolute error and distance of NAM points from their corresponding weather station as follows:

$$DWAE = \frac{1}{n}\sum_{i=1}^{n}\left|\text{NAM\_wind\_speed}_i - \text{station\_wind\_speed}_i\right| \cdot d$$

where:

- NAM_wind_speed_i represents the NAM model's wind speed at a specific NAM point and date i.
- $\text{station\_wind\_speed}_i$ represents the observed wind speed at the corresponding weather station for date i.
- d represents the Haversine Distance of the corresponding NAM point from its weather station.

This method serves a similar purpose in measuring discrepancies like the Mean Absolute Error but it emphasizes larger discrepancies for NAM points further away from its respective weather stations, where wind speed uncertainty may be higher.

2.2.4 Haversine Distance

The Haversine formula calculates the great-circle distance between two points on a sphere given their latitude and longitude:

$$d = 2R \cdot \arcsin\left(\sqrt{a}\right)$$

where:

$$a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)$$

with:

- $R = 6371.0$ km is the Earth's radius.
- ϕ_1, ϕ_2 are the latitudes of the two points in radians.
- λ_1, λ_2 are the longitudes of the two points in radians.
- $\Delta\phi = \phi_2 - \phi_1$ is the difference in latitude.
- $\Delta\lambda = \lambda_2 - \lambda_1$ is the difference in longitude.

This formula ensures an accurate calculation of distances between weather stations and NAM points based on their geographical coordinates.

2.3 Predicting Error of NAM Points Outside the VRI Polygon

As previously mentioned, we filtered the NAM points within San Diego County and Southern Orange County for this purpose. Since NAM points outside the VRI polygons do not have an associated weather station, we will predict the error of these NAM points using a LightGBM model. For our predictor features, we will use the following variables: *nam_wind_speed*, *nam_elevation_m*, *station_elevation_m*, *nam_distance_from_station_km*, *month*, and *day_of_year*. The target variable will be the *abs_wind_speed_error*.

Since the NAM points outside the VRI polygons do not have the values for *station_elevation_m* and *nam_distance_from_station_km*, we will use the values from the nearest weather station to the NAM point. We assume that the wind speed recorded by the nearest weather station should be somewhat similar to that of the corresponding NAM point.

We opted to use a model for error prediction rather than directly using the wind speed from the nearest station because relying solely on the nearest station's wind speed may not fully account for the spatial and temporal variations. The LightGBM model can capture these nuances by considering multiple features, such as elevation differences, distances, and temporal variations, and thus provide a more accurate estimate of the error.

In addition to the spatial features, we also include the *month* and *day_of_year* as temporal features. These allow the model to account for seasonal and diurnal patterns in wind speed variation, which are often significant in weather-related predictions. By directly predicting the error, we can better capture the variability in wind speed predictions and improve the model's ability to generalize across different regions, particularly where direct measurements from weather stations are not available.

2.4 Wind Speed Discrepancy Analysis

The wind speed discrepancy analysis has two key parts. First, we identify NAM points within VRI polygons with high wind speed discrepancies and assess error distribution to locate areas of consistent deviation. Second, we analyze high-error NAM points outside VRI polygons, drawing manual boundaries around them to highlight potential sites for additional wind data collection. By distinguishing errors within and outside VRI polygons, this analysis improves wind speed predictions and informs future data collection.

2.4.1 NAM Points Within VRI Polygons

The first step in analyzing NAM points within VRI polygons involves grouping the data by NAM points and calculating the Mean Absolute Error (MAE) for each point using the MAE formula from Section 2.2.2. This step eliminates temporal variations, allowing us to evaluate the spatial distribution of errors independently. A key aspect of this analysis involves determining the correlation between the *nam_distance_from_station_km*, *station_elevation_m*, *nam_elevation_m*, *abs_elevation_difference* and *abs_wind_speed_error*. This will be achieved by creating a correlation matrix. To better visualize the NAM points, we

will plot it on a Folium map alongside the the weather stations and VRI polgons to identify spatial patterns and trends of the error. Additional analysis includes identifying NAM points with the highest MAE, identifying outlier NAM points, computing the Distance-Weighted Error, and determining which VRI polygons exhibit the highest mean errors. In addition to individual NAM points, we will also investigate the correlation between polygon area size and wind speed error by using a correlation matrix and Folium plot. These analyses will provide insight into how the error varies in space and help to improve future wind speed predictions.

2.4.2 NAM Points Outside VRI Polygon

Analyzing the NAM points outside VRI polygons involves using the predicted absolute error obtained from the LightGBM model. Similar to the previous step, the data is grouped by the NAM points, and the predicted MAE is calculated to remove temporal variations. The NAM points are then plotted on a map to visually examine spatial trends in high-error regions. Since our focus lies on NAM points with a high predicted error, we will be filtering NAM points based on an outlier threshold value. Manual polygon boundaries will be drawn around these points based on the spatial location of the NAM points. Afterwards, we will be plotting SDG&E assets within these boundaries to assess the number of assets and customers located within the areas. This will help SDG&E assess potential locations for additional wind speed data collection and further improve the overall accuracy of wind speed predictions.

2.5 Visualizing Wind Speed Discrepancies

To better understand the discrepancies between NAM wind speeds and weather station wind speeds, we visualize the errors on a map. This visualization overlays the NAM points, weather station points, VRI polygons and conductor spans using Folium. The error intensity is represented through a color gradient for NAM points.

The map includes the following elements:

- **Green Points**: Represent weather station locations.
- **NAM Points (Color-Coded by Error Intensity)**:
  - **Yellow**: Low discrepancy between NAM and weather station wind speeds.
  - **Orange**: Moderate discrepancy.
  - **Red**: High discrepancy.
- **Blue Polygons**: Represent the boundaries of each VRI polygon.
- **Purple Polygons**: Represent the manually drawn polygon.
- **Black Lines**: Represent the conductor spans.

This visualization enables a spatial analysis of NAM wind speed errors, helping to identify regions with significant deviations and potential model improvements.

2.6 Identifying the Nearest Alternative VRI Polygons

To determine the nearest VRI polygons for reassignment, we applied a **Haversine distance-based approach** to measure the geographic distance between NAM points and surrounding VRI polygon centroids. The process involved the following steps:

1. **Extracting Current Assignments** – Each NAM point was first mapped to its existing VRI polygon, recording its corresponding wind speed error.
2. **Finding Alternative VRI Polygons** – For each NAM point, we identified the nearest alternative VRI polygon, *excluding its current assignment*, to ensure we only considered new potential placements.
3. **Reassigning NAM Points** – The NAM points were reassigned to the nearest VRI polygon where the expected wind speed accuracy was improved based on the polygon's wind speed characteristics.

2.6.1 Wind Speed Error Comparison

To evaluate the effectiveness of the reassignments, we compared the original wind speed errors to the errors after the reassignment. The key evaluation metrics included:

- **Mean Absolute Wind Speed Error (MAE) before and after reassignment** – to quantify overall improvement in error reduction.
- **Error difference** – measuring the absolute reduction in wind speed error.

2.6.2 Assessing the Benefit of Reassignment

For NAM points that experienced a reduction in MAE error, we reassigned them to the alternative VRI polygons. To assess the effectiveness of this reassignment, we calculated the following metrics:

- **Overall MAE decrease** – measuring the total reduction in MAE for points that required polygon reassignment.
- **Overall percentage decrease** – evaluating the percentage improvement in prediction accuracy due to reassignment.

3 Results

3.1 Exploratory Data Analysis

Our exploratory data analysis reveals that there are 221 unique weather stations, 308 unique VRI polygons, and over 29,000 recorded wind speed measurements spanning from 2012 to 2024 across 179 unique dates. Similarly, the NAM dataset consists of more than 15,000,000 unique rows over the same 179 dates. There are several columns in the weather

station and VRI polygon data that we can ignore because they are not relevant to our analysis. We also found that both the weather station wind speed and the NAM wind speed distributions are right-skewed, centered at 24 mph (see Figure 1) and 14 mph (see Figure 2) respectively. Furthermore, analyzing the wind speed trends over time reveals fluctuations in both the weather station (see Figure 3) and NAM wind speeds (see Figure 4), highlighting the importance of considering temporal variations. Since the NAM wind speed dataset includes points from across the United States, we filtered for points within San Diego County and Southern Orange County, and plotted it within one histogram (see Figure 5). This comparison shows that NAM wind speeds tend to underestimate the wind speed recorded by the weather stations. In addition, we also computed a correlation matrix between the weather station wind speeds and NAM wind speeds (see Figure 6), which yielded a correlation score of 0.58, indicating a moderate positive correlation. This indicates that NAM wind speed estimates tend to rise as weather station wind speeds increase, though the relationship is not strictly linear.

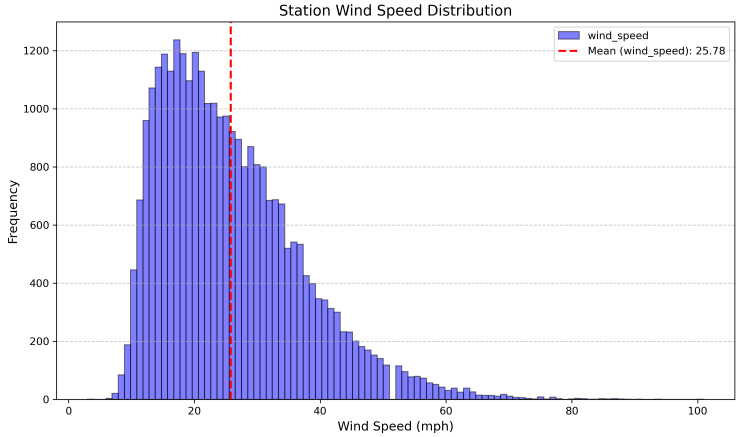

Figure 1: Station Wind Speed Distribution

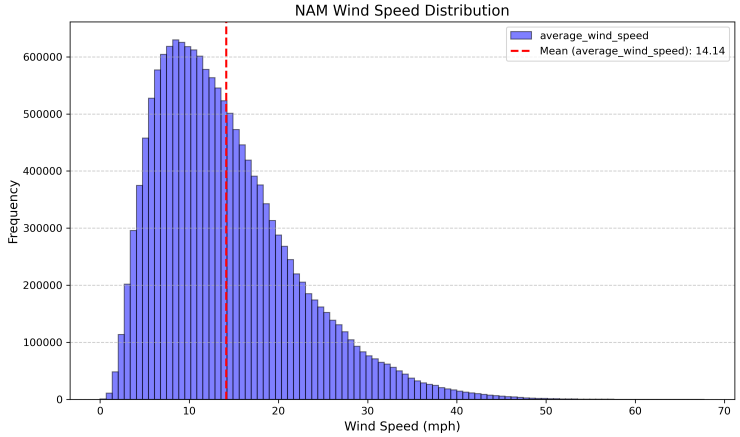

Figure 2: NAM Wind Speed Distribution

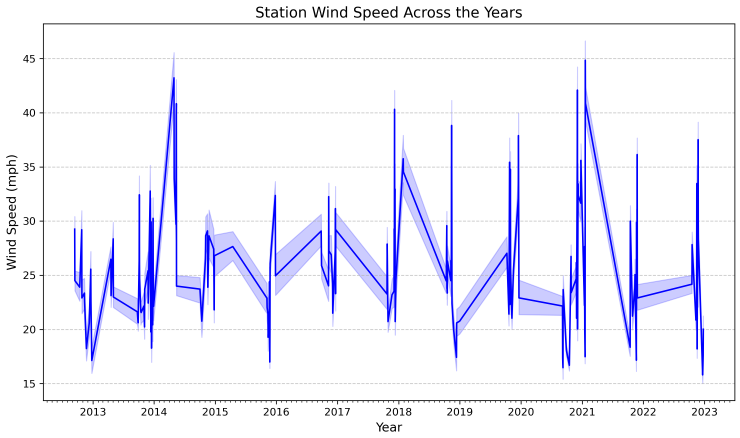

Figure 3: Station Wind Speed Across the Years

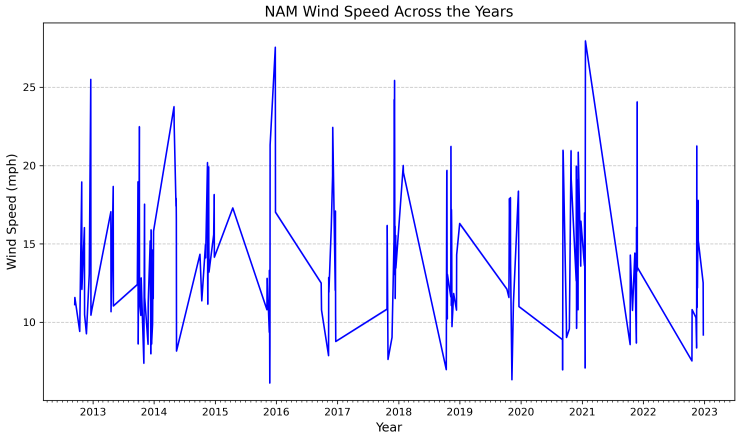

Figure 4: NAM Wind Speed Across the Years

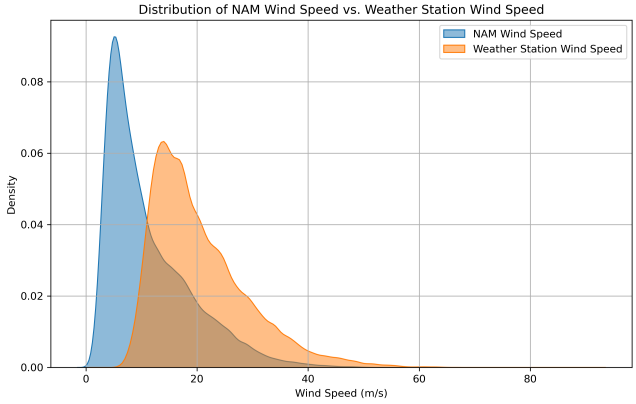

Figure 5: Station & NAM Wind Speed Distribution

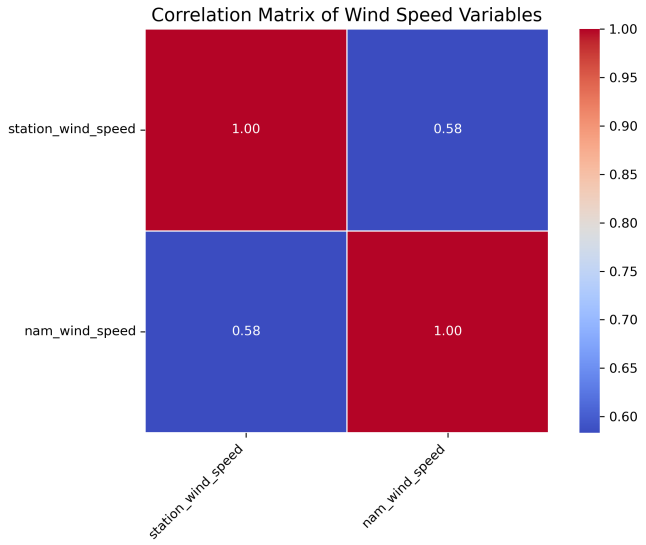

Figure 6: Station & NAM Correlation Matrix

3.2 LightGBM Model

The LightGBM model is trained using NAM points within the VRI polygon which is filtered on corresponding dates. The resulting dataset consist of more than 102,000 data points. This data is split into training and testing sets using an 80:20 ratio. The trained model achieves a mean absolute error (MAE) of approximately 2.7 and an R^2 score of 0.66. We also plot the model's feature importance to identify which factors have the greatest influence (see Figure 7). NAM wind speed is the most significant contributor, with an importance score of 0.3. The trained model is then used to predict the absolute error of NAM points outside the VRI polygons, covering more than 600,000 data points.

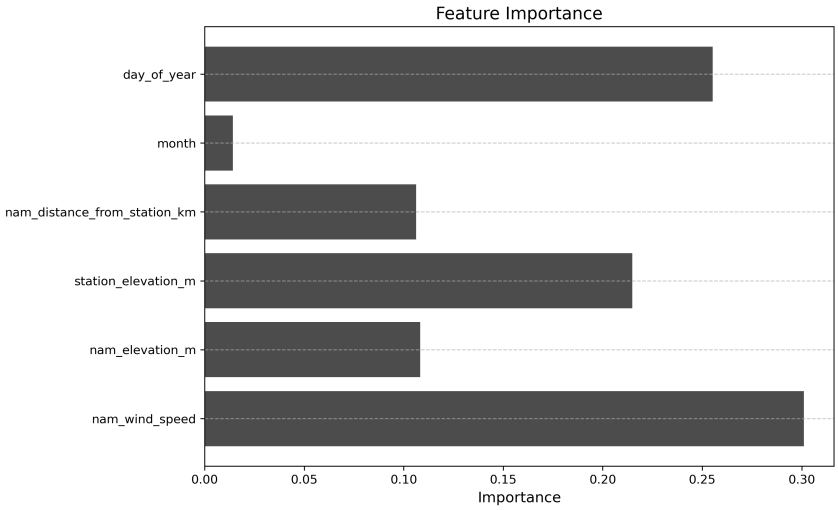

Figure 7: Feature Importance

3.3 NAM Points Within VRI Polygon

The distribution of the Mean Absolute Error of NAM points within the VRI polygon is right-skewed, with a central tendency around 10 mph (see Figure 8), and outlier value of 18.3 mph. The correlation matrix between the Mean Absolute Error and different factors indicate that elevation has a higher positive correlation compared to distance (see Figure 9).

We also generated several map visualizations to spatially analyze the NAM points. The NAM points were filtered based on different criteria, including the outlier threshold value (Figure 10), the top 20 points with the highest Mean Absolute Error (Figure 11), the top 20 points with the highest Distance-Weighted Absolute Error (Figure 12), and NAM points within the top 20 VRI polygons with the highest average Mean Absolute Error (Figure 13).

From these visualizations, we observed that NAM points within thin and elongated VRI polygons, such as TL50003_VRI, tend to exhibit high Mean Absolute Error. Additionally, VRI polygons with high average Mean Absolute Error generally correspond to larger area sizes.

The overall correlation between the average Mean Absolute Error of all VRI polygons and their area size is -0.15, indicating a weak negative relationship. However, when focusing on the 20 polygons with the highest average MAE, the correlation shifts to 0.24, suggesting a weak positive relationship. This implies that, while larger polygons generally exhibit slightly lower errors, the most error-prone polygons tend to show greater discrepancies in NAM wind speed predictions as their size increases.

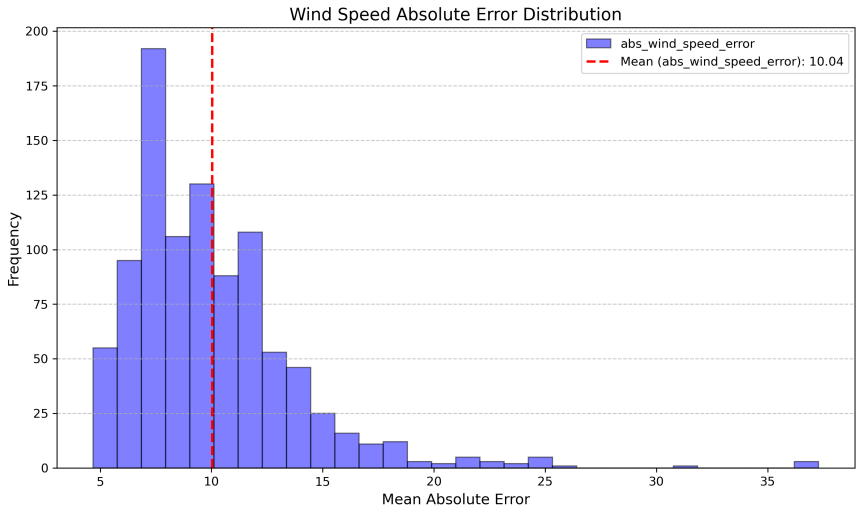

Figure 8: Absolute Wind Speed Error Distribution

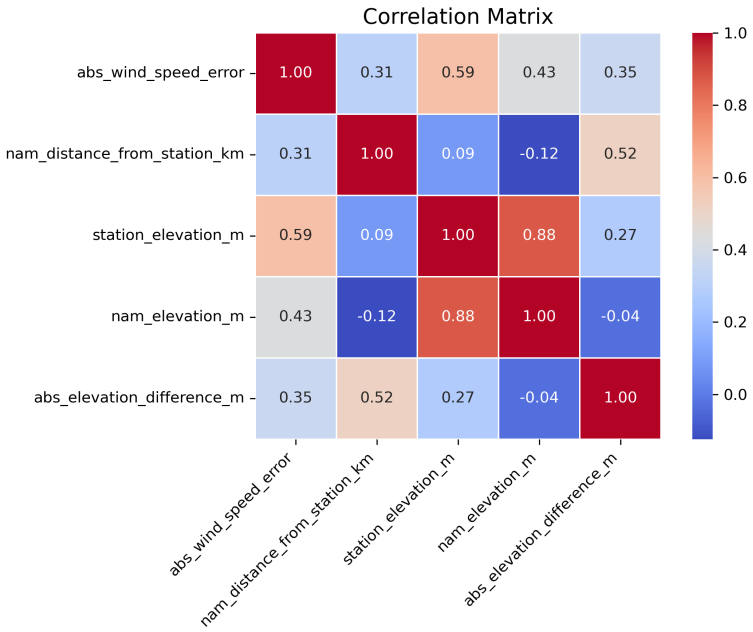

Figure 9: Absolute Wind Speed Error Correlation Matrix

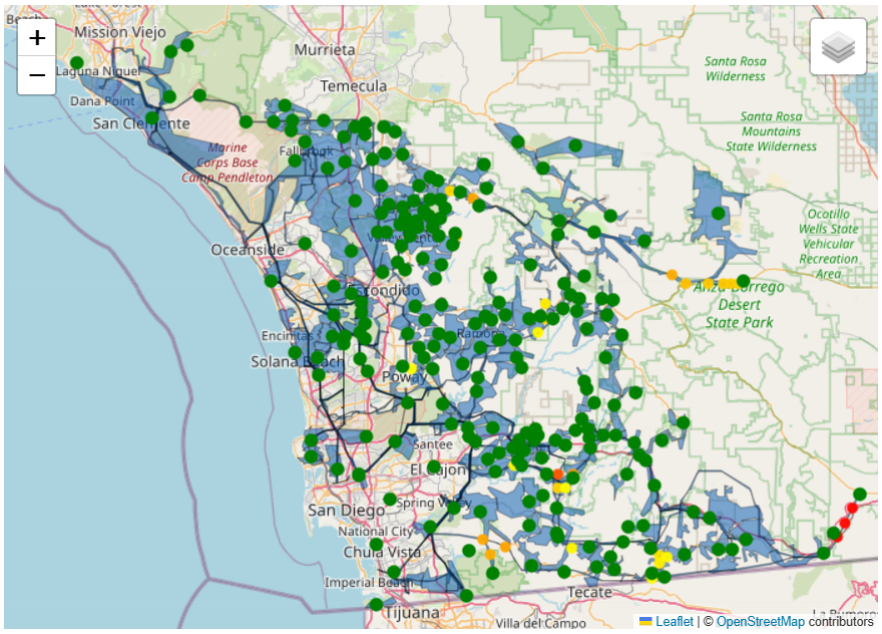

Figure 10: Outlier Wind Speed Error

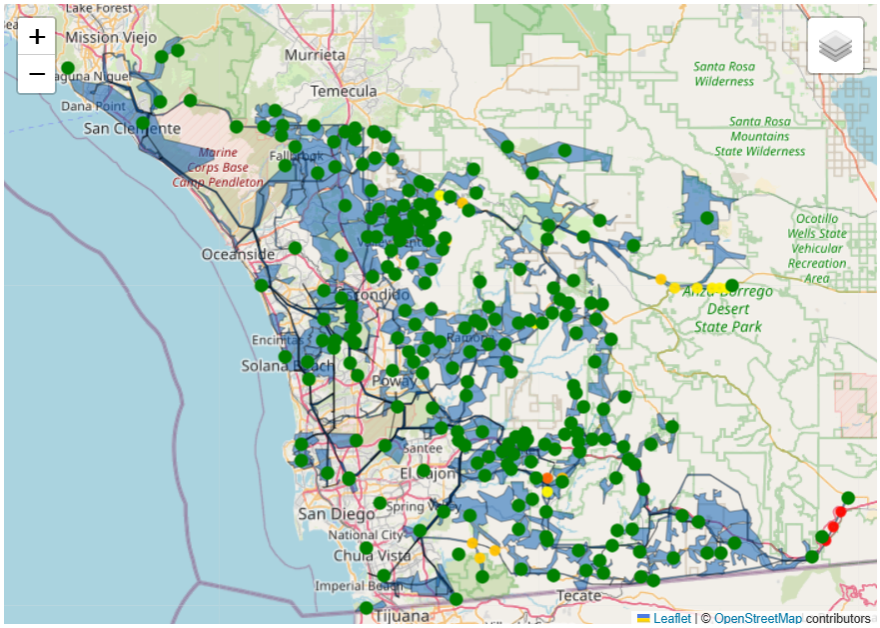

Figure 11: Top 20 Wind Speed Error

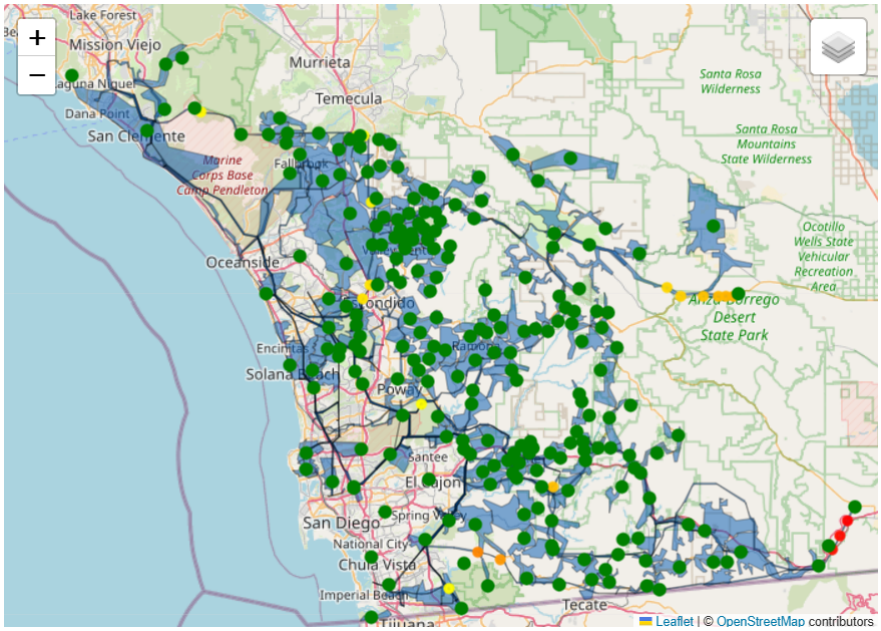

Figure 12: Top 20 Distance-Weighted Wind Speed Error

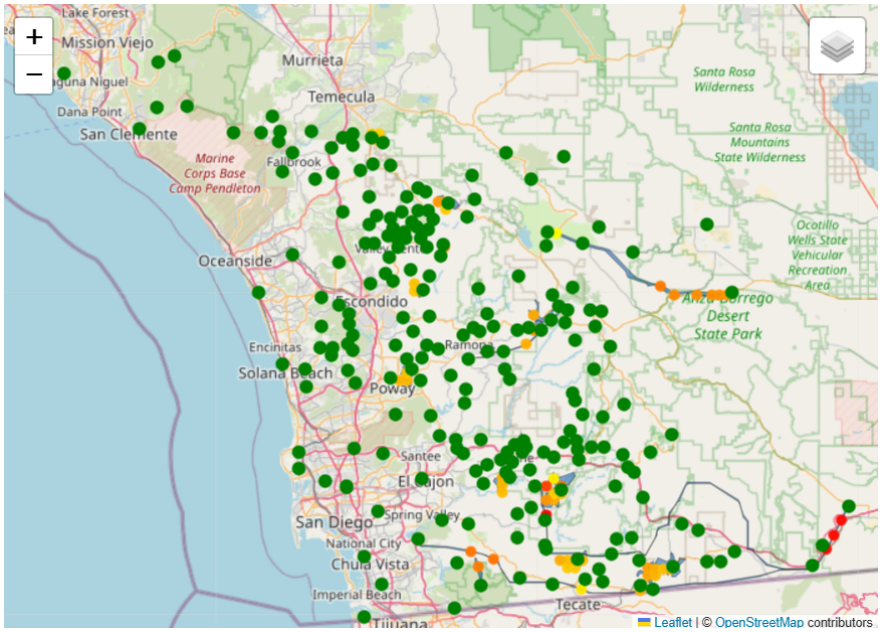

Figure 13: Top 20 Polygon Average Error

3.4 NAM Points Outside VRI Polygon

We generated a correlation matrix between the predicted Mean Absolute Error (MAE) and various factors to identify those contributing most to the error. The results show that station elevation has the strongest correlation (0.90), followed by NAM elevation (0.55) and the distance to the nearest station (0.13). As observed previously, this highlights that elevation plays a significant role in determining wind speed errors.

We also created several map visualizations to analyze spatial trends in wind speed prediction errors. Figure 14 displays all NAM points outside the VRI polygons within SDG&E's area of operation, while Figure 15 highlights NAM points with a predicted Mean Absolute Error exceeding the outlier threshold of 20.096 mph.

From these high-error points, we identified five distinct clusters: Otay Mountains, Cleveland National Forest, CA 78, Sawtooth Mountains, and Monkey Hill. To determine their relevance to SDG&E's operations, we overlaid SDG&E's conductor spans within these areas. Among the five locations, only CA 78 and Sawtooth Mountains contain SDG&E assets and customers. Otay Mesa, despite being within SDG&E's service area, is excluded since it already falls within an existing VRI polygon.

Figure 16 illustrates the boundaries and assets within Sawtooth Mountains and CA 78, while Table 1 provides a detailed count of assets and customers within these regions.

Table 1: Assets & Customers Affected

Boundary	Asset Count	Customer Count
CA 78	459	232
Sawtooth Mountains	401	118

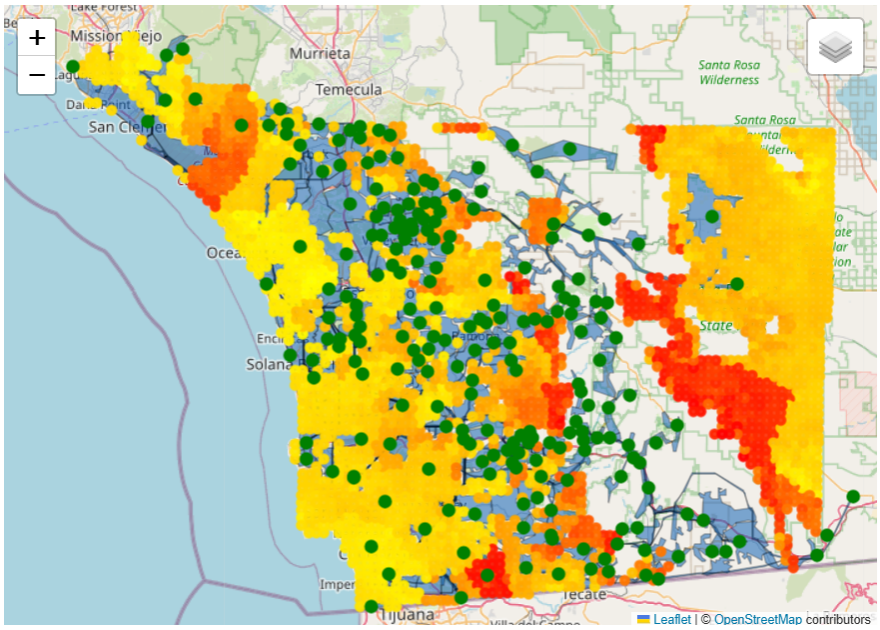

Figure 14: NAM Points Outside VRI Polygon

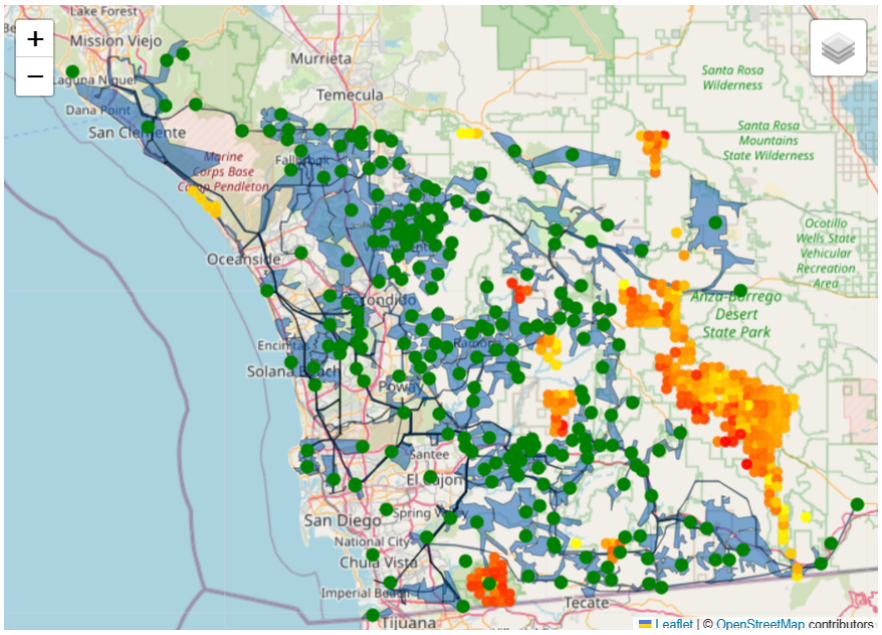

Figure 15: NAM Points Outside VRI Polygon Outlier Error

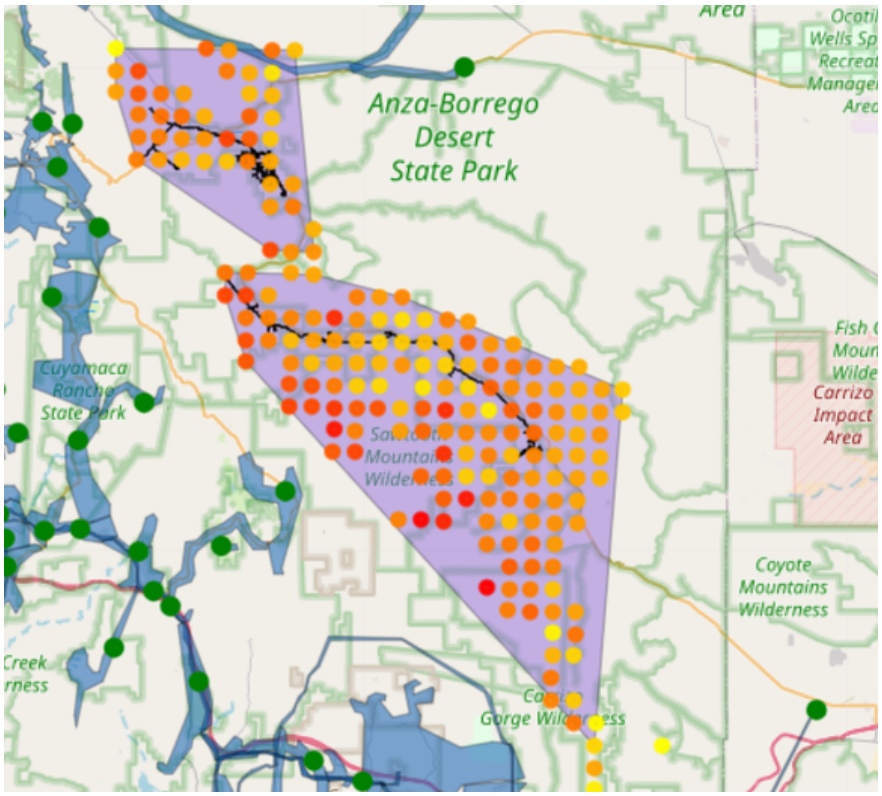

Figure 16: Sawtooth Mountains & CA 78 Boundaries

3.5 Reassigning NAM Points to Nearest VRI Polygon

The results of the VRI polygon reassignment demonstrate a significant reduction in wind speed errors for **18/20** of the top NAM points with the highest MAE errors. For **18/20** of those NAM points, we decided to move on with the reassignment of the VRI polygons in which the remaining NAM points remain assigned to its original VRI polygon since the MAE did not improve. By reallocating NAM points to alternative VRI polygons with better representative wind speed characteristics, the overall prediction accuracy improved. The key findings of the **18 points** in which the MAE is reduced are summarized below:

- The mean absolute wind speed error (MAE) was reduced from **25.76 m/s** to **11.68 m/s**, indicating an overall improvement.
- The average error reduction across reassigned points was **14.07 m/s**.
- The overall percentage decrease in MAE for reassigned points was **54.64%**.

The table below presents a summary of the error reductions for the reassigned NAM points:

Table 2: Summary of Wind Speed Error Reductions After VRI Polygon Reassignment

Metric	Before Reassignment	After Reassignment	Improvement
Mean MAE Wind Speed (m/s)	25.78	11.68	14.07
Overall Percentage Decrease (%)	-	-	54.64%

The reassignment process demonstrates the potential of optimizing VRI polygon allocation to improve wind speed estimations at specific NAM points. Future improvements may involve refining the reassignment criteria and considering additional environmental factors affecting wind speed accuracy.

4 Discussion

4.1 Limited Subset of Dates

As previously mentioned, our project operates on a subset of wind speed data consisting of 179 unique dates. However, this introduces an inherent bias due to imbalances in data distribution across different weather stations. This bias arises because certain weather stations record wind speed observations at significantly different frequencies, often due to differences in their establishment dates and operational periods. For instance, the weather station at North Escondido has only 14 recorded data points, whereas the station at Shockey Truck Trail has the full 179 data points. As a result, the Mean Absolute Error (MAE) of NAM points may not fully capture the true error distribution.

A potential strategy to mitigate this bias is expanding the dataset by incorporating a larger subset of unique dates. By doing so, we can aim to ensure that each weather station has a sufficiently large number of wind speed observations. Additionally, filtering the dataset to include only dates where all weather stations have recorded data could help normalize data representation across locations. However, this approach presents trade-offs: while it enhances fairness in comparison, it may significantly reduce the overall dataset size, potentially limiting the statistical power of our model.

Another alternative is to apply statistical weighting techniques, such as inverse frequency weighting, to balance the contribution of each weather station in the error analysis. This would account for discrepancies in data availability while preserving the full dataset. Future work could explore these methods to refine error estimates and improve the robustness of NAM wind speed predictions.

4.2 Representing NAM Points

The NAM model generates wind speed data at evenly spaced intervals of 1 km, forming a structured grid of points. Our project analyzes these points individually, computing the error for each one. However, this approach presents a limitation, particularly in smaller VRI polygons. In some cases, these polygons may not contain any NAM points simply because the grid points do not happen to fall within their boundaries. This absence of data can lead to gaps in our analysis, making it difficult to accurately assess wind speed prediction errors within these areas.

A potential solution is to calculate errors in a grid-based format rather than at individual points. Instead of analyzing NAM points separately, we can aggregate data from four ad-

jacent points and compute the error for each resulting grid square. This approach offers several advantages. First, it ensures that every region, including small VRI polygons, has an associated error measurement by interpolating values across a structured grid. Second, it provides a more spatially continuous representation of error, reducing the impact of isolated anomalies.

4.3 Manual Boundaries Over Clustering

After predicting the errors of NAM points that do not fall within a VRI polygon, our next goal was to identify distinct polygon boundaries based on the NAM points with significant predicted errors. We attempted statistical grouping methods such as K-means clustering to identify clusters in order to form these boundaries. Although K-means clustering could potentially identify these clusters by grouping NAM points together based on features such as longitude, latitude and errors, we opted to draw the manual polygon boundaries instead to identify these clusters over performing K-means clustering.

One of the key downsides of using K-means clustering in this context is its assumption of spherical clusters of equal variance, which may not accurately reflect the irregular distribution of wind speed errors across different geographical regions. Additionally, K-means requires predefining the number of clusters, which may not always correspond to meaningful physical or environmental divisions. Furthermore, K-means does not account for external geographical constraints, such as terrain variations or existing administrative boundaries, which can be crucial when determining VRI polygons. Drawing manual boundaries allows better flexibility in determining the appropriate boundaries. Moreover, the clusters we are working with is only 5 which makes it practical to draw the boundaries manually in great detail where we can incorporate expert knowledge and ensure that the identified boundaries are appropriate.

4.4 Defining Boundaries for VRI Polygon Optimization

The VRI polygon optimization process aims to enhance the accuracy of NAM point assignments with high errors by reassigning them to more representative VRI polygons. The optimization does not encompass physically reshaping VRI polygon boundaries because various external conditions—such as terrain features, political boundaries, city limits, and regulatory factors, play an important role in defining these polygons. Instead, our results provide valuable insights to geologists and decision-makers at SDG&E by offering data-driven recommendations on potential improvements when drawing VRI polygon boundaries by incorporating our NAM point reassignment suggestions.

4.5 Future Iterations

The results of this project provide valuable insights into areas and VRI polygons with significant wind speed discrepancies. This helps identify regions where additional wind speed

data could improve accuracy. By pinpointing these locations with high error, we can guide the refinement of VRI polygon boundaries to better represent wind speed patterns within each area.

This approach ensures that SDG&E utilizes the most accurate assumption of wind speed data when making decisions about PSPS (Public Safety Power Shutoff) events. By strategically targeting areas with high uncertainty, SDG&E can allocate resources and evaluate these areas more effectively, ultimately minimizing the impact of shutoffs. This targeted approach will also improve grid reliability and contribute to more effective wildfire risk mitigation, leading to a safer and more resilient energy infrastructure.

5 Conclusion

Throughout this project, we have analyzed various challenges related to wind speed inaccuracies involving weather stations, VRI polygons, and NAM points. The first challenge we tackled was addressing wind speed discrepancies between NAM point wind speeds within a VRI polygon and the associated weather station wind speeds. We visualized these discrepancies and conducted an analysis to identify which VRI polygons require particular attention where we observed that NAM points within thin and elongated VRI polygons, such as TL50003_VRI, tend to exhibit high Mean Absolute Error and VRI polygons with high average Mean Absolute Error tend to have larger area sizes.

The second challenge involved addressing NAM points that are not associated with a weather station and are not located within a VRI polygon. To address this, we predicted the errors of those NAM points using a LightGBM model and manually outlined areas with high discrepancies where we identified five distinct clusters: Otay Mountains, Cleveland National Forest, CA 78, Sawtooth Mountains, and Monkey Hill. Notably only CA 78 and Sawtooth Mountains contain SDG&E assets and customers. With the information obtained in regards to identifying these clusters, we narrow down SDG&E search for potential locations for additional wind speed data collection.

For the final challenge, we optimized VRI polygon boundaries by identifying VRI polygons containing the top 20 NAM points with the highest error. Through identifying the nearest alternative VRI polygons of these NAM points and re-assigning the NAM points that experiences error reduction, we achieved a 54.64% reduction in the mean absolute error for the reassigned points. This reassignment demonstrates that wind speed prediction accuracy can be significantly improved by optimizing VRI polygon assignments.

Overall, the methodology implemented in this project provides a framework for addressing wind speed inaccuracies to support SDG&E's infrastructure development plans. Future work, such as expanding the dataset, including additional weather variables, using structured NAM grid areas instead of NAM points, and drawing new boundaries based on VRI polygon reassignments, could further enhance the analysis. These improvements would contribute to more precise wind speed assessments, ultimately strengthening SDG&E's ability to mitigate wildfire risks and enhance grid reliability.

References

- **Zhang, Hailing, Zhaoxia Pu, and Xuebo Zhang**. 2013. "Examination of Errors in Near-Surface Temperature and Wind from WRF Numerical Simulations in Regions of Complex Terrain." *Weather and Forecasting*, **28**(3): 893-914. [Link].

Appendices

A.1 Project Proposal

A.1.1 Abstract

In Quarter 1, our project explored one of the wildfire mitigation strategies involving Public Safety Power Shutoffs (PSPS) by calculating the PSPS probability values. Each value is attributed to a weather station that is operated by SDG&E. Each weather station is associated with a VRI polygon, representing an area where the PSPS probability is uniformly determined by that weather station. This is utilized to find the conductor spans that intersect the polygons and assign each span with a weather station probability. These weather station probabilities are then used to calculate a new span probability which takes into account all the upstream weather station from a span. This highlights the importance of minimizing the error from these weather stations and VRI polygons. Building on the insights gained in Quarter 1, we propose an expanded Quarter 2 project with two primary objectives. First, we aim to identify grid areas with a large error between the weather station wind speed data and the North American Mesoscale (NAM) wind speed data. These areas would be considered as areas of recommendation for new potential weather stations. This would help SDG&E potentially obtain more accurate wind speed data. Second, we aim to refine the VRI polygon boundaries by minimizing a loss functions using the same dataset as the first part of the project. This would be done by implementing an optimization algorithms like gradient descent or heuristic methods, such as simulated annealing or genetic algorithms which would iteratively reduce the loss function. The newly optimized polygon boundaries from the Quarter 2 project can serve as a secondary perspective in the Quarter 1 project, providing a more nuanced evaluation of how PSPS probabilities can be generalized across a given area.

A.1.2 Broad Problem Statement

Wildfires has become an increasing threat to communities, ecosystems, and infrastructure due to various factors ranging from climate change to human activities. Since 2007, SDG&E has installed infrastructures such as weather stations to strengthen precautionary and mitigatory measures against wildfires. Having a precise weather data, particularly wind speed data, is crucial in evaluating wildfire risks and plays a big role in the implementation of mitigation strategies, such as Public Safety Power Shutoffs (PSPS). The risk associated with PSPS is quantified using the PSPS probability which is calculated by dividing number of wind speed data crossing a set alert threshold speed and the count of wind speed data for

each weather station. Each PSPS probability from a weather station is then associated with a VRI polygon, where every area in the polygon have a uniform PSPS probability and identical wind speed data. This highlights the importance of having sufficient weather stations and accurate VRI polygon boundaries since other structures on the electrical grid such as conductor spans are high reliant on this data. This would be especially true with larger VRI polygons which might face accuracy challenges in areas which are located further away from the weather station.

Our project tackles these challenges by identifying grid areas for potential placements of new weather stations in areas with high discrepency between the VRI polygon wind speed measurements and wind speed data from the North American Mesoscale (NAM) model. In addition, we plan to modify the VRI polygon boundaries by minimizing the error between the weather station data and the NAM model data. This will be achieved by developing an error function that quantifies these discrepancies by using both the NAM data and weather station measurements which would allow us to pinpoint areas with larger error measurements. Furthermore, by developing an optimization algorithm, new VRI boundaries that minimize a loss function can be drawn which will provide a more accurate area representation of VRI polygon, computed based on historical data of wind speeds. This would lead to an more consistent assessment of the probability values assigned to the VRI polygons.

The goals outlined in our project are both impactful and achievable within a 10-week timeframe. By addressing potential improvements in weather monitoring accuracy and quantifying VRI boundaries, this project aims to provide a recommendation on areas to build new weather stations to enable increased accuracy of real data collection and improving the VRI boundary to optimize the generalization of PSPS probability within a set area. Having accurate weather data is essential for preemptive measures such as resource allocation, strategic planning, and deciding actions like the Public Safety Power Shutoffs. This project will lay the groundwork for more efficient and data-driven wildfire risk management strategies for future potential projects.

A.1.3 Narrow Problem Statement

The Quarter 1 project involves working with 5 different datasets containing weather stations, weather station alert speed threshold, wind speed data, VRI polygons and conductor span dataset. The first part of the project involves working with the first 3 datasets aforementioned to calculate the PSPS probability of each weather station. Using the PSPS probability of each weather station alongside the remaining 2 datasets, the PSPS probability of each span and estimated annual customers affected within a segment/circuit can be calculated. All the data used in the project was given to us and was produced by different teams within the company. For example, the VRI polygon boundaries were decided by the meteorologists based on several different factors. Instead of building on the results from the Quarter 1 project, our Quarter 2 project focuses on validating and assessing the accuracy of the existing data by incorporating a new comparative wind speed dataset from the North American Mesoscale (NAM) model. This will be done by minimizing the error between the existing wind speed data and the new wind speed data. By doing so, this will enable us to

identify areas with high error margins to recommend new placements of weather stations as well as utilize an optimization algorithm to recommend new VRI polygon boundaries.

To reiterate our goal, we aim to deliver two recommendations: (1) Identifying areas with a high error score and recommend potential areas to install new weather stations, and (2) Redefining VRI polygon boundaries by minimizing a loss function between the actual weather station wind speed data and the NAM model data. This will not replace the existing framework but will serve as a secondary perspective. It can be integrated into the Quarter 1 project to analyze how the PSPS probability of the conductor span changes when using the updated polygon boundaries as well as predicted number of customers affected.

To accomplish the first task, we will conduct a time-series analysis using wind speed data from weather stations and the NAM model for corresponding dates. Each point in the NAM dataset represents a 1 km^2 area, and for each point intersecting a VRI polygon, we will calculate the error between the NAM wind speed and weather station measurements using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Bias Error (MBE). To ensure accuracy, we will preprocess the data by aligning timestamps, filtering NAM data to include only relevant points, and normalizing the datasets. Temporal trends will be analyzed using rolling windows to capture seasonal variations and error fluctuations during critical high-wind events. Spatial error patterns will be visualized by overlaying the NAM geometry with the computed error values on a map which provides a clear representation of discrepancies across regions. Time-series plots will also be made to highlight trends and variations in these discrepancies over time which enables a comprehensive understanding of temporal and spatial dynamics. Finally, sensitivity analyses and validation checks will ensure consistency and assess the impact of NAM resolution on the results by applying varying weights to NAM data points based on their distance to weather stations and using statistical correlation metrics, such as Pearson or Spearman coefficients, to evaluate the relationship between NAM predictions and observed weather station data.

To complete the second task, we will use the same dataset to iteratively adjust the boundaries of the VRI polygons and propose new optimized boundaries. This involves calculating a loss function for every targeted polygon and continuously refining its shape to minimize the loss. The loss function will measure the error between NAM wind speed predictions and weather station measurements within each VRI polygon, incorporating metrics such as Mean Absolute Error (MAE) or Root Mean Square Error (RMSE), and weighted by factors like proximity to weather stations within the polygon. Geometric transformations, such as scaling and vertex movement, will be applied iteratively to adjust the VRI polygon boundaries. Optimization algorithms like gradient descent or heuristic methods, such as simulated annealing or genetic algorithms, will be used to iteratively reduce the loss function. To ensure meaningful results, convergence criteria will include determining thresholds for the loss function and constraints to prevent unreasonable boundary shifts, such as maintaining geographic continuity or avoiding overlaps with other polygons. Once optimized, the proposed boundaries will be validated using a holdout dataset to confirm improvements in accuracy without overfitting. Comparisons between the updated and original polygons will be made using error metrics and visual analyses. This would be visualized by overlaying the original and updated boundaries on a map alongside a heatmap to highlight regions

with the most significant accuracy improvements. This iterative approach aims to minimize prediction errors and enhance the reliability of the span PSPS probability calculation.

Overall, this project aims to identify optimal locations for additional weather station placements and redefine the existing VRI boundaries. Building upon the foundation established in the Quarter 1 project, we seek to explore how our findings may alter the results from that initial analysis. This new perspective can offer valuable insights when making decisions about wildfire mitigation strategies and resource allocation, potentially enhancing the effectiveness of efforts to protect communities and infrastructure from wildfire risks. By integrating updated data and refining the VRI polygons, we aim to improve the accuracy of PSPS probability predictions, leading to better-informed strategies for managing wildfire-related risks.

A.1.4 Primary Output Statement

This project will have two major output with potentially 2 deliverables. The first output will consist of a set of 1 km^2 square geometries, each accompanied by their respective error values, highlighting areas with high levels of error. These regions will serve as the proposed locations for new weather stations. The error distribution will be visualized on a map using Folium, with a gradient color scale to illustrate the intensity of the error, overlaid with the original VRI polygons. The second output will include a set of newly optimized polygon boundaries designed to minimize the error between weather station data and NAM data. These updated boundaries will be displayed on a map alongside the original VRI polygons, allowing for a visual comparison to recognize the changes made. Additionally, the results from the first task will be incorporated to assess whether the recommended locations for new weather stations still apply when using the updated VRI boundaries. These maps will be incorporated as static images in the final report and may also be hosted on a server which allows for interactive exploration of the data through a website.

A.1.5 Data Source Statement

Given that this project aims to optimize the data used in Quarter 1, it will involve working with datasets provided by SDG&E to build upon the foundation established in the first project. Specifically, we will utilize the existing weather station data, wind speed data, and VRI polygon data from the Quarter 1 project.

In addition to these datasets, we will incorporate a new source of wind speed data from the North American Mesoscale (NAM) model. This open-source dataset is available here. The NAM wind speed data will serve as a comparative dataset to assess discrepancies between the existing weather station data and the NAM model, helping us to calculate the error and refine the model's accuracy. After consulting with our mentors, we have confirmed the accessibility and validity of the NAM dataset.

The NAM dataset consists of raster data, which is a type of geographic data that represents information as a grid of pixels. Each pixel corresponds to a specific location on Earth and

holds a value that represents a particular attribute, such as wind speed. To integrate this dataset with the VRI polygon data, we will convert the raster format into a geospatial data type compatible with the polygon representation, enabling us to perform the necessary comparisons and optimizations for the project.

A.2 Modifications to original proposal in Q2

Through implementing our project, we discussed our approaches to tackling the two objectives that we introduced in our initial proposal mentioned involving:

- Identifying wind speed discrepancy between NAM points and weather station points
- Optimizing VRI polygons by minimizing a loss function through an algorithm

However, our mentor advised us to focus on the first objective in our project. Hence, the second objective would be more of a supplementary objective to the first objective. Our brand new direction for the supplementary objective would be to utilize the insights gained from identifying which NAM points have a high discrepancy and using those insights to perform analysis on how those VRI polygons associated with those NAM points with high discrepancy could be optimized.

6. Contribution Statement

This project was a collaborative effort among all team members, each contributing to different aspects of the research, data analysis, and reporting. Below is a breakdown of the contributions made by each team member:

- **Kenneth Hidayat**: Explored the initial datasets, researched how .nc file works and how to convert it to a panda dataframe, helped with preprocessing the data, contributed to the EDA of NAM wind speed data and weather station data patterns. Additionally, contributed to the implementation of error functions and the final report write-up. Kenneth also performed optimization of VRI polygons and designed the poster for the final deliverable.
- **Steven Sahar**: Focused on the geospatial analysis of VRI polygons and weather station data. Developed visualizations for our analysis using Folium to highlight discrepancies in wind speed measurements. Also implemented the model to predict the discrepancy errors of NAM points not within a VRI polygon. Lastly, Steven developed the website and contributed to the report for the final deliverable

Each team member was actively involved in discussions, feedback sessions, and iterative improvements to ensure the project's success. This collaborative effort enabled a comprehensive approach to identifying wind speed discrepancies and optimizing VRI polygon boundaries for wildfire mitigation strategies.