Project 4 (Final Project) Proposal – Using Multilinear Regression of Publicly-Provided NAM Forecast Data to Develop Site-Specific Wind and Temperature Models

Group members: Steph Abegg

The research question

I work for LongPath Technologies, a Boulder-based company that has created revolutionary laser-based technology to monitor methane gas emissions. We monitor emissions at hundreds of sites across the United States. The technology works by measuring the methane concentration on either side of the site, and using the difference in measurements along with wind speed and wind direction data to compute plume models for the methane on the site. So site-specific accurate wind measurements are vital to accurate emission readings. LongPath has a three-dimensional (3D) anemometer installed at each site to collect wind data.

One weakness of 3D anemometers is that they are susceptible to icing up. LongPath has several sites in places like North Dakota, Wyoming, and Colorado, where winter storms cause the anemometers to ice up for days on end. Without wind data, it is not possible to compute the emissions from the measured methane concentrations, so the system is essentially down. It is usually the case that the anemometer is the only component of the system that is not.

A question is whether publicly-provided forecast data, such as from the North American Mesoscale (NAM) Forecast System, can be used as a proxy for the anemometer data during the times when the anemometer is down. To address this question, I compare 3D anemometer data to NAM data for the same location during the same 30-day period of time. The data from the 3D anemometer is recorded on 5-second intervals. The data is first averaged over 15-minute windows (this smooths out the data as well as corresponds to how the wind data is used in practice), and then the 15-minute averaged temperatures, wind directions, and wind speeds are directly compared to the NAM forecast data for these same time windows using timeseries plots, regression analysis, and binning.

The next question—and the meat of the analysis in this study—is whether a multilinear regression model based on the NAM data can provide even better site-specific predictions of the wind and temperature conditions than the direct NAM forecasts alone. To address this question, I use NAM forecast data to train site-specific multilinear regression models for temperature, wind direction, and wind speed.

The data

This study uses two datasets.

The first dataset is from a 3D anemometer located North Dakota at 47.8437 N, 102.8524 W, elevation 2300 ft above sea level. The data spans 30 days from February 11, 2024 to March 11, 2024. The anemometer measures on five-second intervals. There are two two-day gaps in the data, corresponding to when the anemometer was iced up: February 22 and 23 and March 3 and 4. So there are a total of 26 days of anemometer data. The raw 3D anemometer data contains 420,917 rows (pared down to 420,911 rows after cleaning the data).

The relevant columns include:

- Date and time in UTC;
- Number of internal data points used to compute the measurements corresponding to a single time;
- Temperature in degrees Celsius;
- Wind direction in degrees (North: 0°, East: 90°);
- Wind speed in meters per second;
- Wind elevation in degrees.

In practice the anemometer data is averaged only 15-minute intervals. After doing so, there were 2344 rows of data.

The second dataset is NAM forecast data. The NAM forecast data can be requested for ftp download from the government website. Pulling out the relevant forecast data required writing a python script. The script looped through hundreds of the downloaded .grb2 files, extracting the desired information (i.e. temperature, horizontal and vertical components of wind velocity, vertical velocity, and kinetic energy) at the correct atmospheric level (i.e. surface) and at the correct location (i.e. 47.8437 N, 102.8524 W). The forecast periods of interest were narrowed down to 0, 1, 2, 3, 4, 6, 12, 24, 48, 72 hours. Finally, the extracted NAM forecast data was saved into a .csv file. The NAM forecast data used in this study contains 1318 data points (i.e. four times daily for 33 days, for all 10 forecasting periods).

The NAM dataset has numerous columns, but the ones extracted for this analysis are:

- Date and time in UTC;
- Forecast period in hours (0, 1, 2, 3, 4, 6, 12, 24, 48, 72 hours);
- Temperature in degrees Celsius (surface level);
- Wind direction in degrees (planetaryBoundaryLayer:level, 80 m);
- Wind speed in meters per second (planetaryBoundaryLayer:level, 80 m);
- Vertical velocity, m/s (isobaricInPa:level 100,000 Pa) (used to compute wind elevation);
- Kinetic energy in J/kg (isobaricInPa:level 100,000 Pa).

Satisfying the project requirements

My project meets the project requirements, detailed a follows:

- My project is about a problem worth solving, analyzing, and visualizing. In fact, my (previous) company plans to implement the results.
- My project will use machine learning, since my project focuses on developing a multilinear regression model (regression is type of supervised learning method) to use forecast data to model site-specific conditions.
- My project will use Scikit-learn for the multilinear regression.
- My project dataset contains well over 100 records.
- From the list of "you must use at least two of the following", my project will use:
    - Python Pandas
    - Python Matplotlib
- Since my project is based on Python analysis, my project will use several other additional python packages/libraries, such as:
    - Python datetime
    - Python numpy
    - Python math
    - Python scipy.stats
    - Python pygrib
    - Python os
    - Python requests
    - Python bs4 BeautifulSoup
    - Python tarfile