Using Multilinear Regression of Publicly-Provided NAM Forecast Data to Develop Site-Specific Wind and Temperature Models

Stephanie Abegg • November-December 2024 • Final Project

Table of Contents

Research questions

This study has two primary research questions.

Is forecast data a good proxy for anemometer data?

The first research question is how well publicly-provided forecast data correlates with anemometer measurements. Temporarily using forecast data instead of measurement data would be beneficial in case an anemometer is malfunctioning, iced up, or even non-existent. It could also be used to predict the movement of the methane gas plumes. To address this question, I analyze data from the North American Mesoscale (NAM) Forecast System, collected for the same 30-day date range as the anemometer data and corresponding to the same location. The NAM forecast data is computed every six hours, for the selected forecast periods of 0, 1, 2, 3, 4, 6, 12, 24, 48, and 72 hours. The NAM forecast data for temperature, wind direction, and wind speed is directly compared to the 15-minute averaged anemometer data via timeseries plots, regression analysis, and binning. The forecast windows are separated to analyze the accuracy of the forecast from zero hours to three days.

Can a multilinear regression model trained on forecast data predict site-specific anemometer data even better?

Next, in an attempt to find better site-specific correlation of the forecast data, the NAM forecast data is used to train site-specific multilinear regression models for temperature, wind direction, and wind speed. These models provide predictions that produce an even better correlation with the measurements at the site of the 3D anemometer.

The Data

The datasets

There are two raw datasets used in this study. These are:

1. Data from a 3D anemometer in North Dakota, located at 47.8437 N, 102.8524 W. The elevation at this location is approximately 2300 ft above sea level and the 3D anemometer is situated 3 meters

above the ground. The 3D anemometer data spans 30 days from February 11, 2024 to March 11, 2024, on five-second intervals. There are two two-day gaps in the data, over the days of February 22 and 23 and March 3 and 4, corresponding to winter storms when the 3D anemometer was iced up. (So, there are a total of 26 days of data for the 3D anemometer, although throughout this paper it is referred to as 30 days due to the time span of the data.) The raw 3D anemometer data contains 420,917 rows (pared down to 420,911 rows after cleaning the data). The relevant columns include:

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

Figure 1. A 3D anemometer.

2. Forecast data for the location of the anemometer over the same date range as the anemometer data. This data is from the North American Mesoscale (NAM) Forecast System 12km grid for all of North America. The NAM Forecast System is one of the major regional weather forecast models run by the National Centers for Environmental Prediction (NCEP) for producing weather forecasts. The data is provided to the public free of charge. Dozens of weather parameters are available from the NAM grids, such as temperature, precipitation, wind speed, wind direction, and turbulent kinetic energy. The 12 km grid forecast is run four times daily at 00z, 06z, 12z and 18z with forecast periods from 0 hours to 84 hours with a 1-hour temporal resolution. Since the 0-hour forecast is the initial state of the model, which is based on weather observations leading up to the forecast predictions, the 0-hour forecast is considered close to "truth".

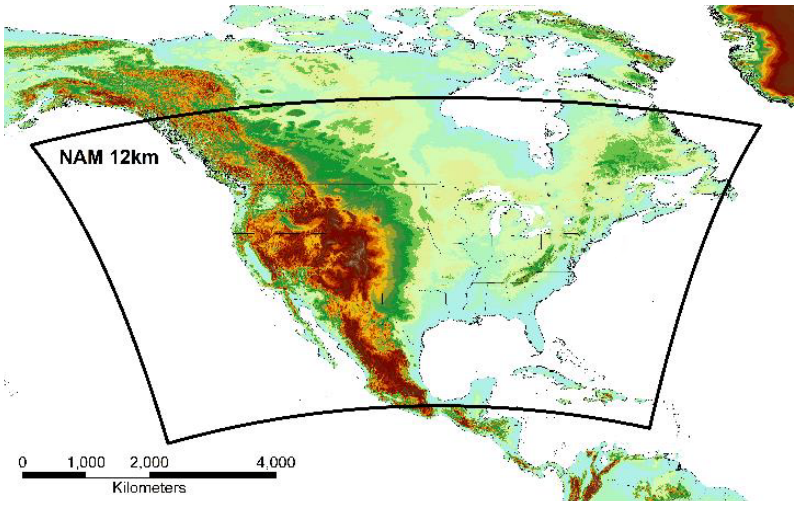

Figure 2. NAM 12 km grid.

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:

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

Cleaning the data

Before analysis could ensue, the datasets had to be cleaned. The dataset cleaning is described below, where "3D" and "NAM" indicate which dataset(s) was involved in each cleaning step.

- NAM: The cleaning took place as the data was being extracted from the individual downloaded .grb2 files. Only data needed for the analysis was included in the final exported .csv file, and the columns were named as desired.
- 3D: The UTC time column was changed into a datetime64 format.
- 3D: The seconds column was added to the UTC time, which curiously had all of the seconds zeroed out in the original dataset.
- 3D: Columns that were not needed for the analysis were removed. Examples of columns that were removed were "_id", "Origin of data", and various pre-calculated date, hour, minute, second, and second bins that were later added back in forms more useful to this analysis.
- 3D: Columns were renamed to clarify units (e.g. "temp" to "temp_C", "wspd" to "wspd_mps").
- 3D: Rows with null values were dropped.
- 3D: The measurements from the anemometer is in five-second intervals. An individual measurement is an internal computation from several data points taken over the preceding five seconds. This number is given in the "n_pts" column. A typical measurement from the 3D anemometer is computed from about 100 individual data points wh. The histograms indicated that some measurements had just a couple of data points contributing to the measurement. Any rows where the number of internal data points was less than five were removed.
- 3D: The data was checked for duplicate rows. There were no duplicate rows.
- 3D: To make the wind speeds from the 3D anemometer comparable with the wind speeds from the NAM forecast data (which was computed upon extraction to be the horizontal wind speed), the wind speed from the 3D anemometer was adjusted to just the horizontal component multiplying by the cosine of the wind elevation. The wind elevation angles are generally small, so this correction had very little effect on the wind speeds.
- NAM: The 3D anemometer and the NAM forecast data corresponded to different heights above the ground. The 3D anemometer is 3 meters above the ground and the elevation of the "surface" wind speed measurements from the NAM data is 80 meters off the ground. To be comparable, the wind speeds were corrected using the wind shear formula ($v_2 = v_1 \frac{(h_2/z_0)}{(h_1/z_0)}$, where v_1 is the reference wind

speed measured at height h_1, v_2 is the wind speed at height h_2, and z_0 is the roughness length which depends on the terrain). The height of the 3D anemometer was used for reference, and the NAM wind speeds were corrected to represent a height of 3 meters above the ground.

Adding columns

After the data was cleaned, some new columns were added. The new columns are described below, where "3D" and "NAM" indicate which dataset(s) was involved in each column addition.

- 3D/NAM: To make the time series plots more representative of a local day (midnight to midnight), a column for local time was added. The location of the anemometer is in Central Time, which is UTC -06:00, so computing local time involved subtracting 6 hours from the UTC time.
- 3D/NAM: Columns were added for date and hour of day, both in local time.
- 3D: A column was added for minute. This was used for binning into 15-minute intervals.
- 3D: A minute binning column was added to bin the 3D anemometer data into 15-minute intervals, starting on the hour (i.e. [0, 15), [15, 30), [30, 45), [45, 60)). These bins were used for averaging the temperature, wind speed, and wind direction over every 15 minutes. This smooths out the data as well as corresponds to how the wind data is used in practice.
- NAM: A column was added for the date and hour of the forecast, in local time. This involved adding the forecast period (in hours) to the local time to get the time corresponding to the forecast, and then separating the result into date and hour. A future dataframe was merged on these new forecast date and hour columns to be able to directly compare the forecasts with each other and with corresponding anemometer measurements for that date and hour.
- 3D/NAM: Columns were added for the cosines and sines of the wind direction. This was needed for the circular averaging of the anemometer wind direction measurements over the 15-minute bins. Circular averaging averages the east-west and north-south components separately, and then computes the average wind angle by finding the arctangent of the ratio of the components. Also, the cosine and sine components of the wind direction were needed for the multilinear regression models. Standard linear regression cannot handle the circular nature of time properly because it assumes a linear relationship and does not account for the circular wrap-around of angles.
- 3D/NAM: Columns were added for the cosines and sines of the scaled hour of day. These were needed for the multilinear regression models, for the same reason as the wind components.
- 3D/NAM: A column was added for temperature in degrees Fahrenheit. Fahrenheit is the unit used for the plots, being more commonplace in the United States than Celsius.
- 3D/NAM: A column was added for wind speed in miles per hour. Miles per hour is the unit used for the plots, being more commonplace in the United States than meters per second.

Figure 3 shows histograms of the date, temperature, wind speed, and wind directions in the 15-minute averaged 3D anemometer data and NAM data. The data ranges generally match up for the anemometer and NAM forecast data.

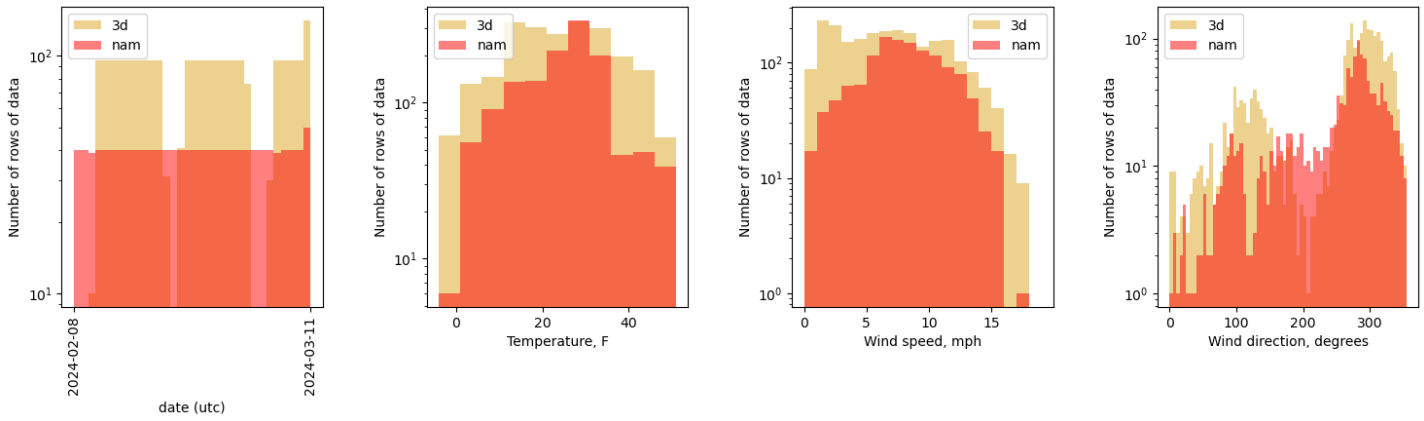

Figure 3. Histograms of the 15-minute averaged 3D anemometer data and the NAM forecast data.

Joined dataframe

Now that the data was cleaned and the necessary columns added, it was time to merge the 3D anemometer data and NAM data to allow for direct comparison of the data for the same times. The analyses and multilinear regression were based primarily on this merged data. A dataframe "df_3D_and_nam" was created by via an outer join of the 15-minute averaged 3D anemometer data and the NAM data, joining on the date and hour of the 3D anemometer data and forecast date and hour of the NAM data. Since the NAM data is every six hours (i.e. four times a day), this means that for a given forecast period, there is a 3D anemometer vs. NAM comparison at a frequency of six hours, with four 3D data points for every NAM data point (because each NAM data point corresponds to an hour, and the 3D data points are for every 15 minutes).

Columns were added for NAM-3D temperature difference, wind direction difference (corrected to be between -180 to 180), and wind speed difference.

Figure 4 shows the number of rows of data in the final joined dataframe.

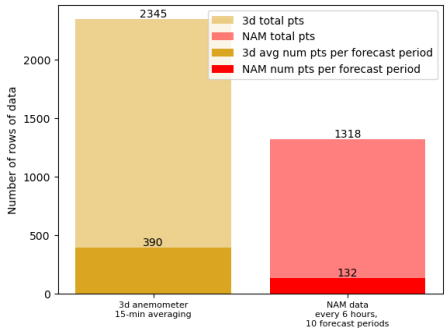

Figure 4. Number of rows of data in the joined dataframe. For any given forecast period, there are about four times as many 3D anemometer measurements (ideally there would be exactly four times as many, but there were a couple of time periods when the anemometer was iced up).

Key assumptions

Here are some key assumptions made in this study:

- The anemometer is oriented correctly and was in normal operation during the timespan of the dataset.
- The anemometer is placed in the open and not located close to any heat sink or source.
- The wind shear formula applies to correcting the wind speeds for height. The NAM "surface" measurement is for a height of 80 meters above the ground, so the wind speeds could follow a slightly different distribution than at the surface.
- The 92,500 Pa level is appropriate for the surface-level measurement for the geometric vertical velocity and turbulent kinetic energy. The elevation of the location is about 2300 feet, which has an estimated pressure of 92,630 Pa at 0°C.
- The anemometer data was collected for 30 days during the winter in North Dakota, when weather can be quite harsh. This analysis assumes that the weather and conditions do not affect the results. It also assumes that this 30-day window is representative of general behavior at any time.

Direct Analysis of NAM Forecast vs. Anemometer Data

Timeseries comparison

Figure 5 shows timeseries of the NAM forecast data and the 15-minute averaged temperature, wind direction, and wind speed measurements of the 3D anemometer over the 30-day span of data. Visually, the NAM forecasts and the anemometer data line up. The difference between the NAM forecast data and the anemometer data (NAM-anemometer) is shown in blue. The differences hovers near zero, but has some scatter.

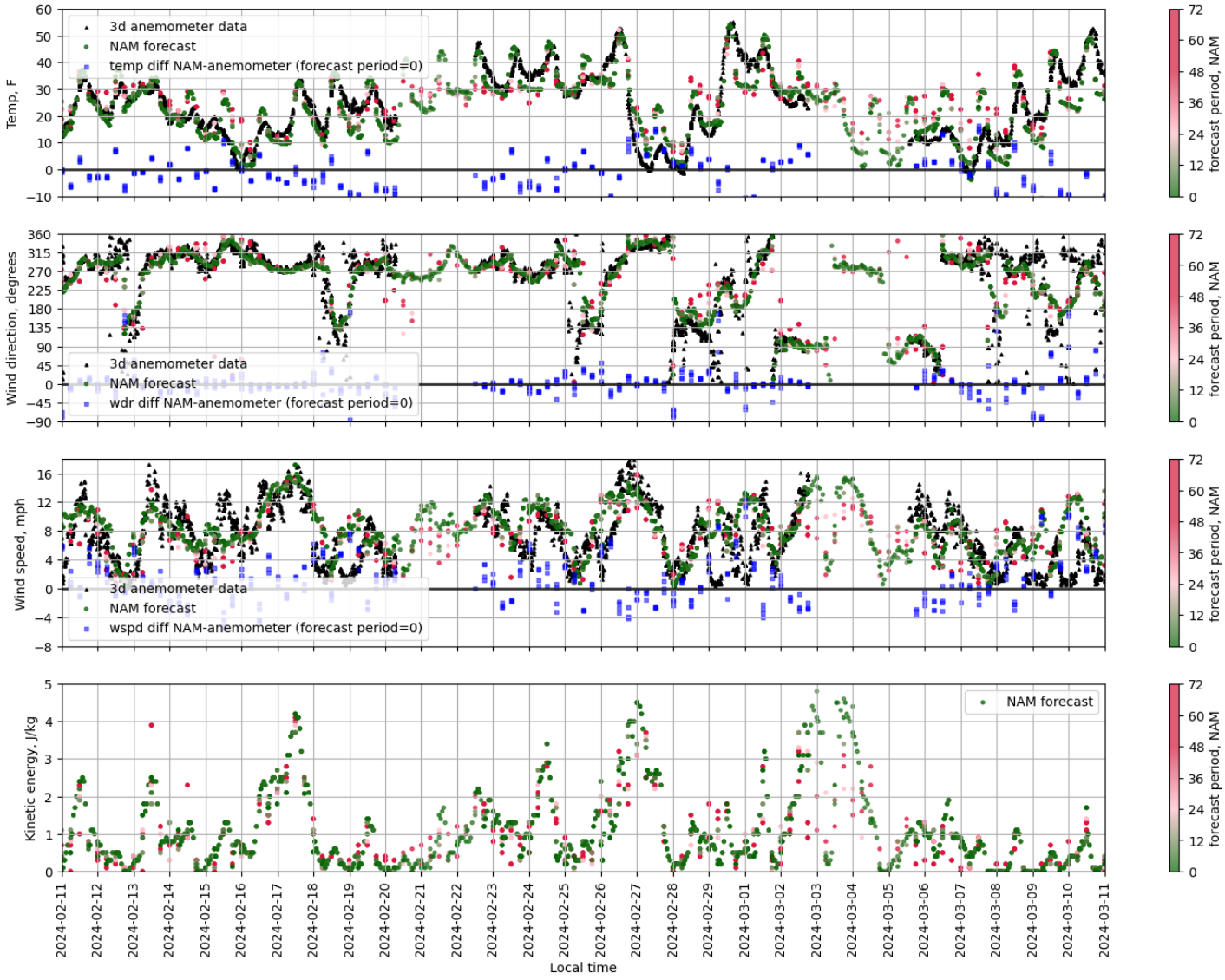

Figure 5. NAM forecast data and 3D anemometer data (15-minute averaged) for temperature, wind direction, and wind speed. The difference (NAM-anemometer) is shown in blue. The NAM forecast data is color-coded by forecast period, where green represents a forecast of half a day or less, and red represents a forecast of up to 1-3 days.

The differences in temperature, wind direction, and wind speed between the NAM forecasts and the 3D anemometer and various influencing factors will be investigated and discussed in following sections.

Since the 0-hour forecast is the initial state of the model, which is based on weather observations leading up to the forecast predictions, the 0-hour forecast is close to “truth”. As expected, the further out forecasts (such as 36-72 hours) do occasionally stray from the shorter-term forecasts (such as 0 to 12 hours). But it seems like the forecast has little variation for any forecast period less than 12 hours, indicating that a 12-hour forecast is sufficient to predict the 0-hour forecast.

Wind speed, wind direction, temperature differences per binned wind direction, wind speed, and kinetic energy

Figure 6 shows histograms of the differences between the NAM forecast data and the 3D anemometer 15-minute averages for the same 15-minute windows. The forecast period is limited to 0-12 hours to remove the inaccuracies of longer-term forecasting and the wind speeds to greater than 1 m/s (2.24 mph) to remove the inaccuracies at low wind speeds. The distributions are centered around 0.

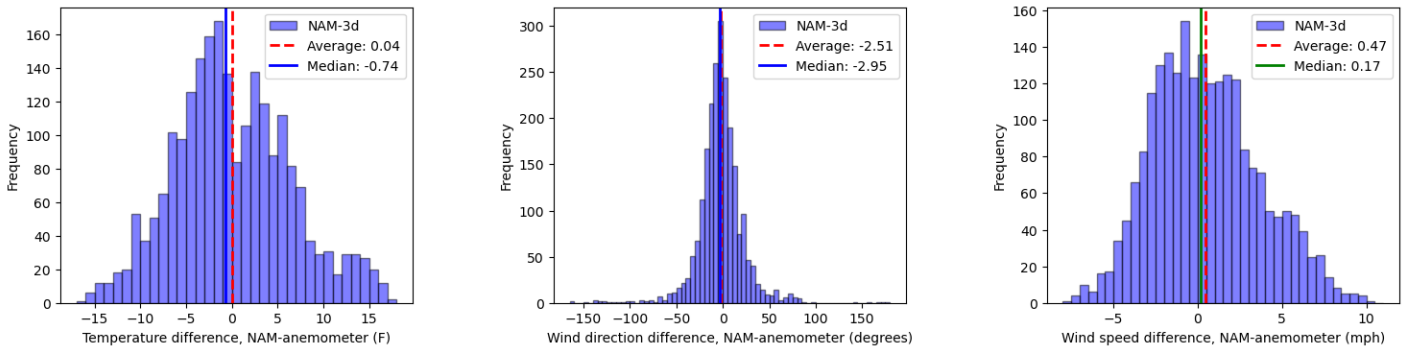

Figure 6. Histograms of the difference between the NAM forecast data and 3D anemometer 15-minute averages for the same 15-minute windows. The forecast period is limited to 0-12 hours and the wind speeds > 1m/s (2.24 mph).

Figure 7 shows the average temperature difference, wind direction difference, and wind speed difference between the NAM forecast and the 3D anemometer (NAM-anemometer) as a function of the (binned) temperature, wind direction, wind speed, and kinetic energy of the NAM forecast. This way of looking at the data investigates whether there are any trends in the differences with various metrics. The data shown is for forecast periods of 0-12 hours.

Over the 30-day span of data, the average temperature difference (NAM-anemometer) between the forecast data and 3D anemometer data was 0.04°F. So, the average NAM temperature forecast is essentially the same as the average anemometer temperature measurements. The top row in *Figure 7* shows that the NAM-anemometer difference follows roughly a sinusoidal pattern with wind direction and wind speed. There is no obvious explanation for this, other than perhaps terrain influences. There is a distinct trend of higher forecasted temperatures for higher turbulent kinetic energy.

Over the 30-day span of data, the average wind direction difference (NAM-anemometer) between the forecast data and 3D anemometer data was about -2.5 degrees. This is well within the 20 degree (or so) tolerance of wind direction measurement of an anemometer. The middle row in *Figure 7* (again!) shows again that the NAM-anemometer difference follows roughly a sinusoidal pattern with wind direction and wind speed. This time, kinetic energy does not seem to have much effect on the difference between the wind direction forecast and measurement.

Over the 30-day span of data, the average wind speed difference (NAM-anemometer) between the forecast data and 3D anemometer data was about 0.47 mph. Again, this is quite small and well within any tolerances of the anemometer for wind speed measurements. The bottom row in *Figure 7* shows (yet again!) that the NAM-anemometer difference follows roughly a sinusoidal pattern with wind direction and wind speed. Also,

we see that at lower temperature and turbulent kinetic energy, the NAM forecast tends to have higher wind speeds than that measured by the anemometer.

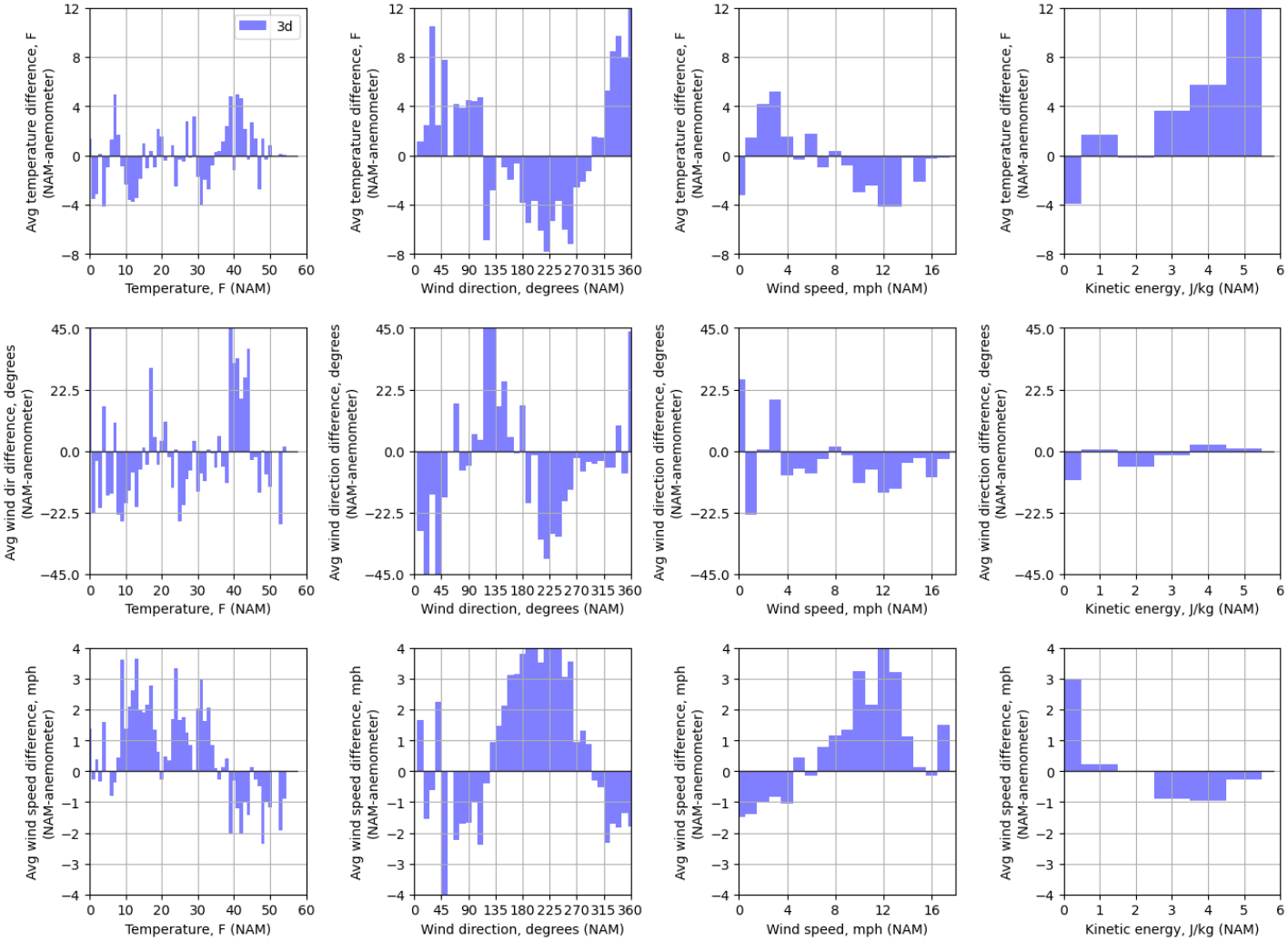

Figure 7. Average temperature difference, wind direction difference, and wind speed difference (NAM-anemometer) per binned temperature, wind direction, wind speed, and kinetic energy of NAM forecast. The forecast period is limited to 0-12 hours to remove inaccuracies from longer-term forecasting.

Regression analysis

The NAM forecast was chosen to be for the same location (i.e. the corresponding segment on a 12-km grid spacing) as the 3D anemometer, so under ideal behavior the forecast would match the anemometer measurements, barring for small differences due to modeled or physical height and terrain. Simple linear regressions were conducted to quantify the linear relationship between the NAM forecasts and the measurements of the anemometer. Separate regressions were done for the ten different forecast periods (0, 1, 2, 3, 4, 6, 12, 24, 48, 72 hours), to see how the linear relationship changed with forecast period.

Figure 8 shows the linear regression between the NAM forecasted temperature and the 15-minute averaged temperature measurements of the 3D anemometer. Each data point corresponds to the same hour of time, where there are four anemometer data points for every NAM datapoint. We see that for all forecast periods, the data follows a linear relationship with correlation coefficients of 0.73-0.89, which indicates a strong positive linear relationship. The lower correlation coefficients are seen for the forecast periods past 24 hours.

Figure 9 shows the linear regression between the NAM forecasted wind direction and the 15-minute averaged wind direction measurements of the 3D anemometer. Each data point corresponds to the same hour of time, where there are four anemometer data points for every NAM datapoint. We see that for all forecast periods, the data follows a linear relationship with correlation coefficients of 0.80-0.87, which indicates a strong positive linear relationship. Interestingly, there is no drop in correlation for greater forecast periods, indicating that wind direction can be forecasted accurately a few days out.

Figure 10 shows the linear regression between the NAM forecasted wind speed and the 15-minute averaged wind speed measurements of the 3D anemometer. Each data point corresponds to the same hour of time, where there are four anemometer data points for every NAM datapoint. We see that for all forecast periods, there is quite a bit of scatter in the data, and the correlation coefficients range from 0.36-0.60, which indicates a moderate positive linear relationship between the forecast and measurements. There is a noticeable association between the forecasted and measured wind speeds, but there is a significant amount of variability. The lower correlation coefficients are seen for the forecast periods past 24 hours.

An important observation is that there is a fair bit of scatter in these NAM versus anemometer regressions.

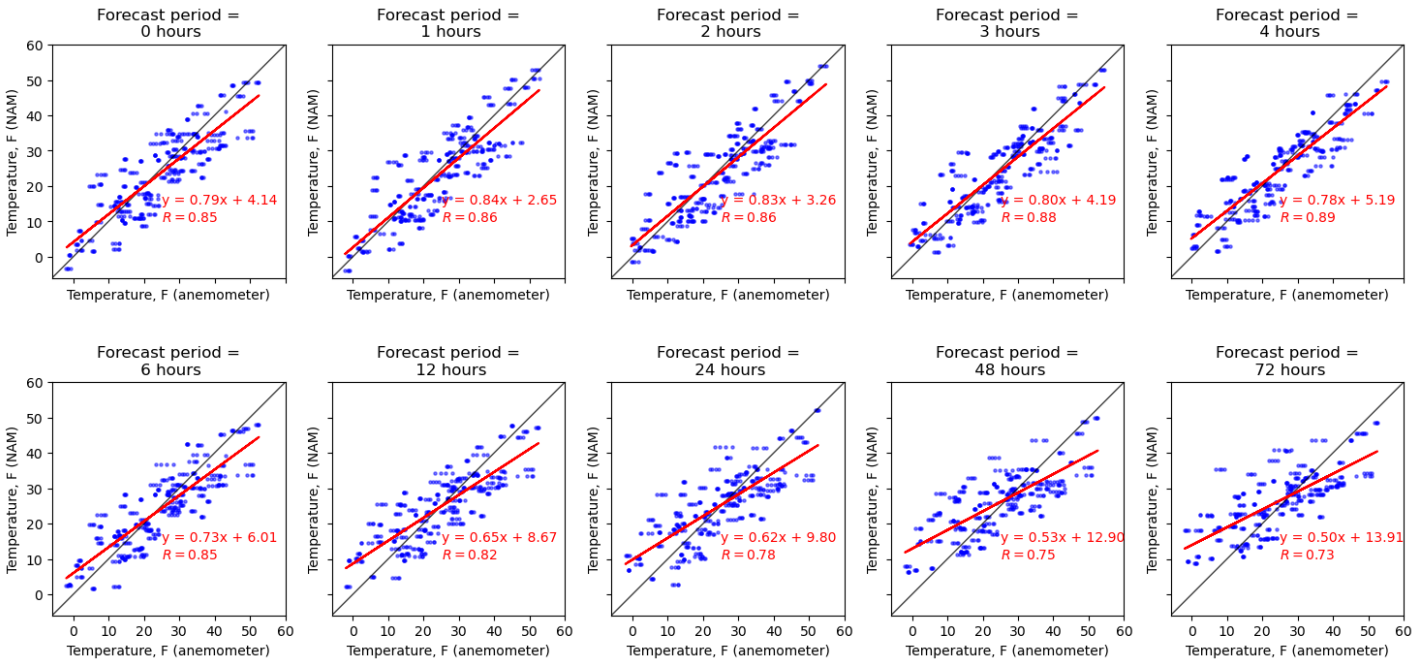

Figure 8. Linear regression of NAM forecasted temperature versus the 3D anemometer temperature measurements.

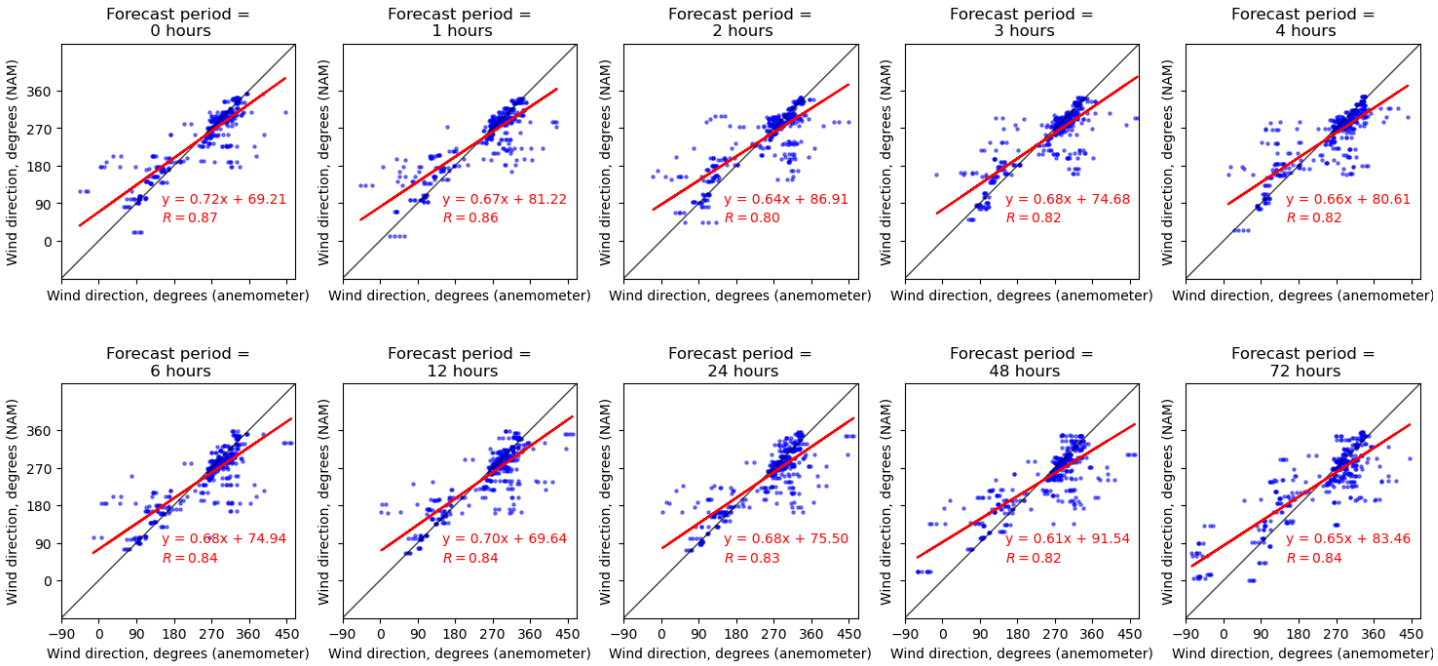

Figure 9. Linear regression of NAM forecasted wind direction versus the 3D anemometer wind direction measurements.

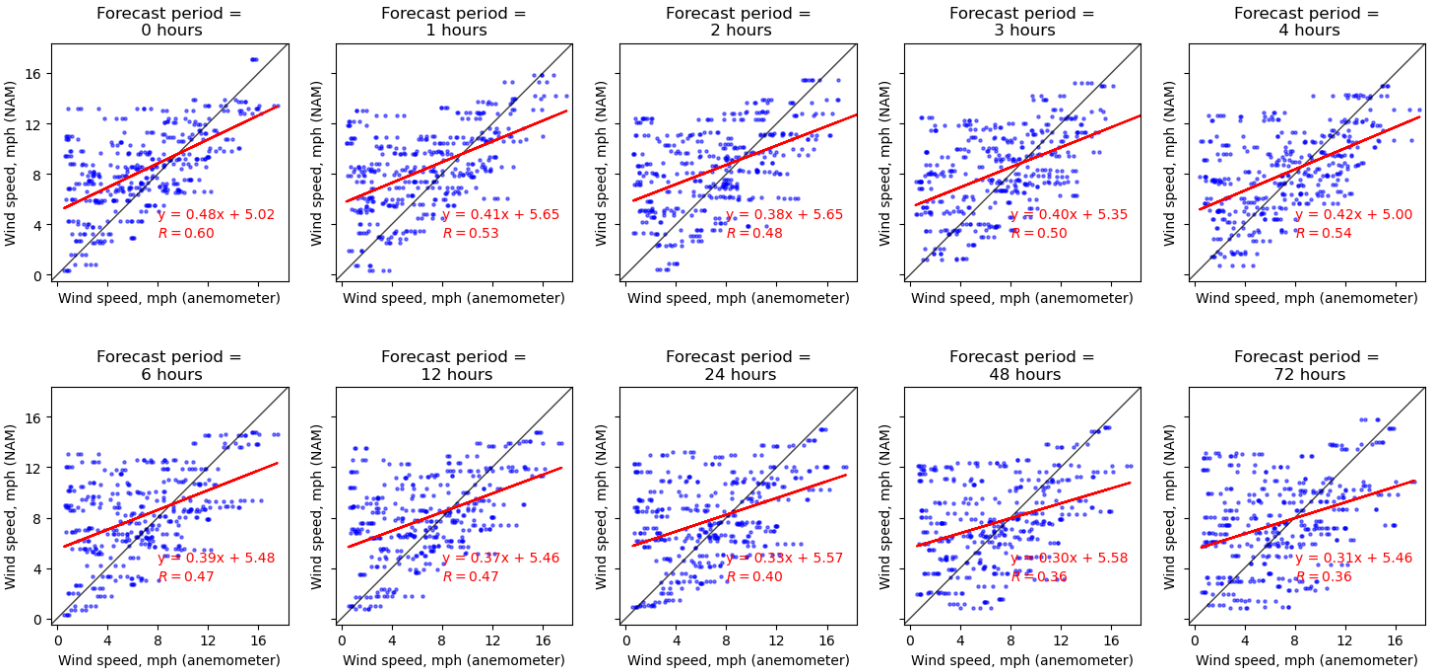

Figure 10. Linear regression of NAM forecasted wind speed versus the 3D anemometer wind speed measurements.

Turbulent kinetic energy

The NAM model provides a forecast for "turbulent kinetic energy." Turbulent kinetic energy refers to the kinetic energy associated with turbulent motion in the atmosphere. Turbulence plays a significant role in mixing air masses, influencing weather patterns, and distributing heat and momentum. Turbulent kinetic energy is a measure of how much energy is contained in the turbulence and is often used in weather models to estimate the intensity of turbulent motions within the atmosphere.

An interesting question if turbulent kinetic energy can serve as a predictor of differences between the NAM forecast and the 3D anemometer measurements. *Figure 11*, *Figure 12*, and *Figure 13* show scatter plots of the difference between the NAM forecasted data and the 15-minute averaged anemometer measurements for temperature, wind direction, and wind speed, respectively. The most notable trend is seen with wind direction, where lower kinetic energies result in a greater difference in NAM forecast and anemometer measurements and higher kinetic energies result in a significantly smaller difference. In other words, the NAM forecast is a better approximation of the measured data when there is higher turbulence. This relationship holds for all forecast periods, but is strongest for shorter term forecasts.

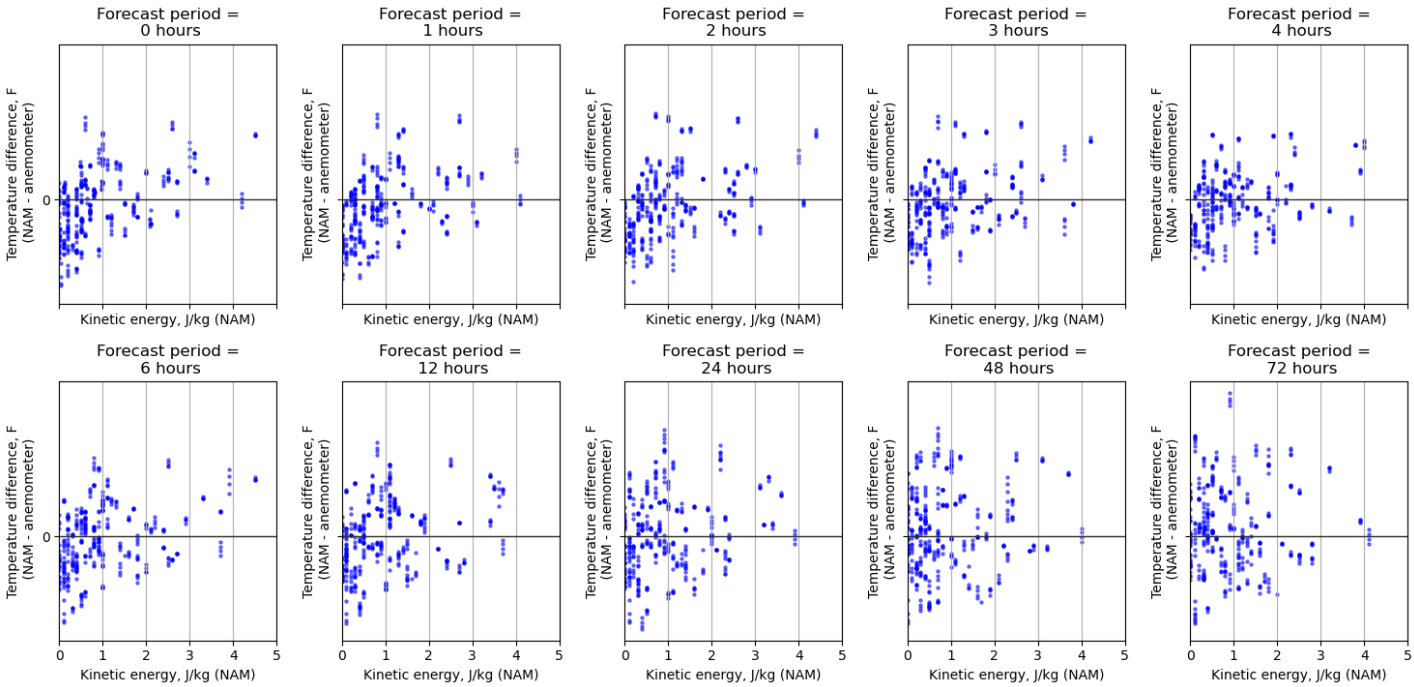

Figure 11. Scatterplots of temperature difference (NAM-anemometer) between forecast and measurements versus forecasted turbulent kinetic energy.

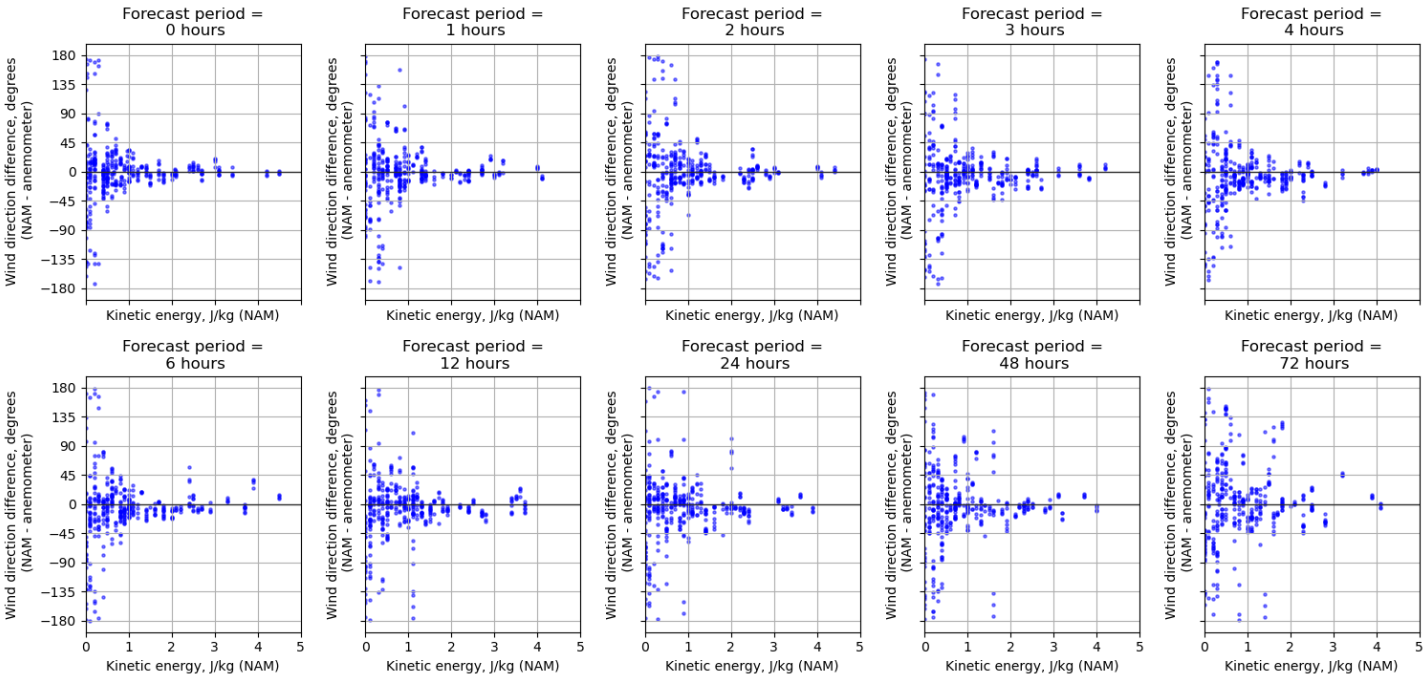

Figure 12. Scatterplots of wind direction difference (NAM-anemometer) between forecast and measurements versus forecasted turbulent kinetic energy.

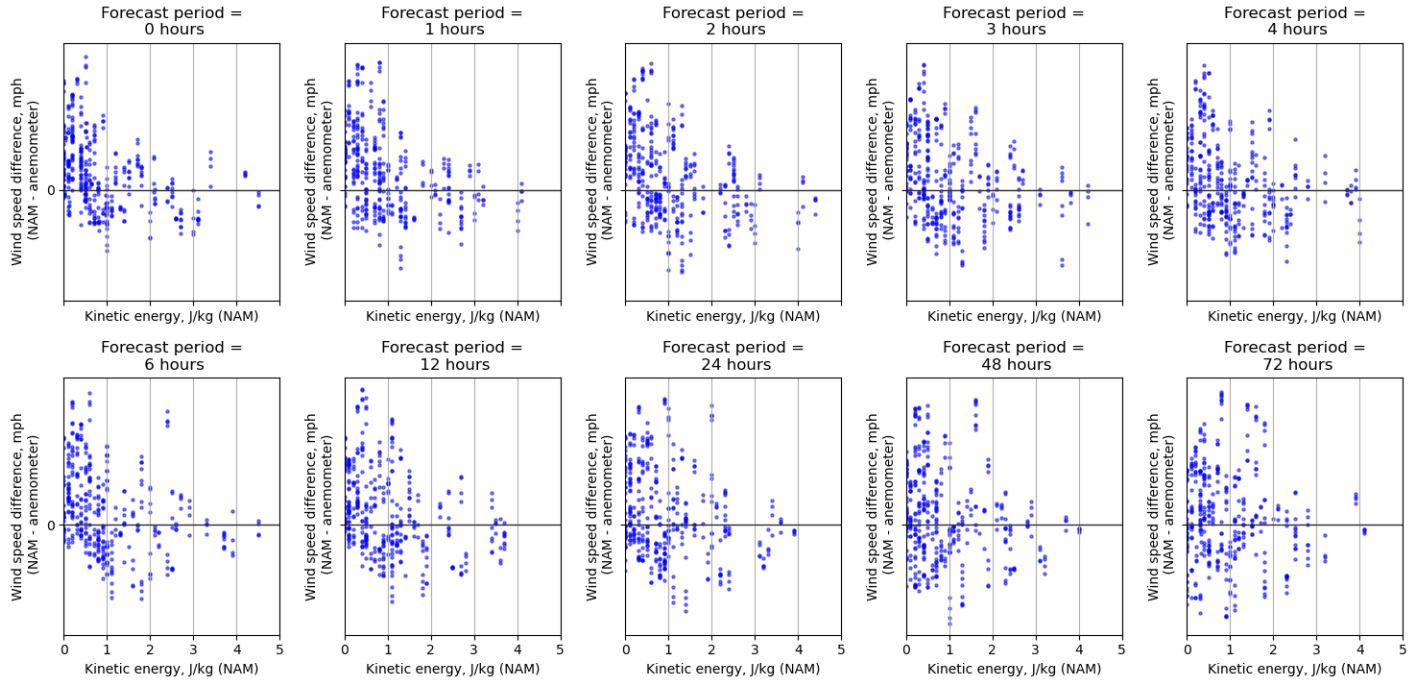

Figure 13. Scatterplots of wind speed difference (NAM-anemometer) between forecast and measurements versus forecasted turbulent kinetic energy.

Multilinear Regression Models

Can a combination of the NAM forecast data offer better predictions?

A primary goal of this study is to determine if NAM forecast data can be used to accurately predict the conditions at the location of the 3D anemometer. Up to this point, we have looked at a direct comparison of the temperature, wind direction, and wind speed NAM forecast values with the respective 15-minute averaged 3D anemometer measurements. However, what if considering more of the NAM forecast variables, including looking backwards and forwards in time, would provide predictions that give an even better approximation of the 3D anemometer measurements?

This is where a multilinear regression model comes into play. This is a regression model where there is one dependent variable (the target) and multiple independent variables (predictors). The model attempts to find a linear relationship between the dependent variable and multiple independent variables. Here we assume that the 3D anemometer presents the truth, and NAM data and temporal elements (such as hour of day and day of year) are the predictors. The dependent variables in question are the temperature, wind direction, and wind speed at the anemometer location. A multilinear regression model was created for each one in turn.

The NAM data used for the model was for forecast periods of 1-12 hours. The first reason for this choice of forecast periods is that 0-12 hours was shown to be a conservative window when forecasting is stable (see the timeseries plots of the NAM data *Figure 5*). The second reason for this choice of forecast periods—namely in using a minimum forecast period of 1 hour —is that the NAM forecast data is released on six hour intervals, so the minimum forecast period available for real-time predictions is anywhere from 1-6 hours, depending on when the prediction is being made; using a minimum forecast period of greater than 1 (such as 6 hours) would limit the number of useable data points too much, so hence 1 hour was chosen, which is still reasonable given the stability of forecasts in the 0-12 forecast hour window. The NAM data used in the model was also filtered to wind speeds of greater than 1m/s (2.24 mph), since these low wind speeds are filtered out in practice, and were shown earlier in this study to be the cause of much of the scatter in the data. This resulted in 1912 data points spanning the 30-day window usable for developing the regression models. The train-test split was 80/20 percent.

Python's Scikit-learn package was used for the multilinear regression. The predictor variables selected from the NAM forecast data were: temperature, cosine and sine of wind direction, wind speed, forecast period, wind elevation, kinetic energy. Additional temporal predictor variables were the cosine and sine of forecast hour and the forecast day of year. Lagging variables, i.e. containing values from previous ("_b#") or subsequent ("_f#") steps in time, were created for temperature, wind direction components, wind speed, and kinetic energy. The data was split into a training set (80% of the data, or 1540 data points) and testing set (the remaining 20% of the data, or 384 data points). The training set is used to train the model while the testing set is used to evaluate the model's performance. Time series data needs to be split while preserving the sequence, so rather than using a random test-train split, the training set was the last 25 days of data and the testing set was the first 5 days of data (this could have been reversed to the first 25 days and last 5 days with similar results).

Figure 14, *Figure 15*, and *Figure 16* show plots of the 3D anemometer data, model predictions, and NAM forecasts, for the three dependent variables (i.e. temperature, wind direction, and wind speed). The plots

represent the test set, which spanned the first five days of data. Notice how in general, the model predictions tend to follow the up and down trends of the 3D anemometer better than the NAM forecast.

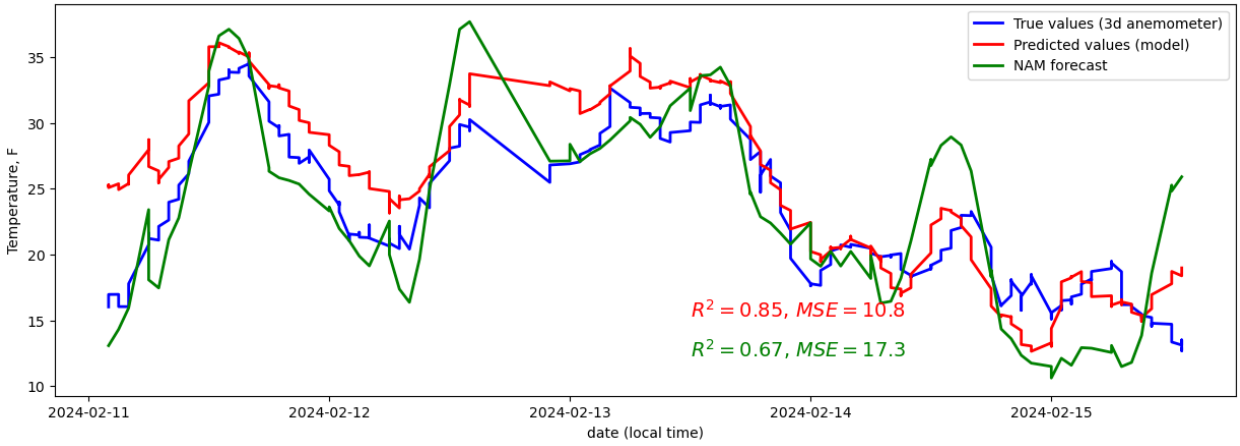

Figure 14. Model values and NAM forecast values and 3D anemometer measurements of temperature. The model has a higher R^2 value and lower MSE than the NAM forecast.

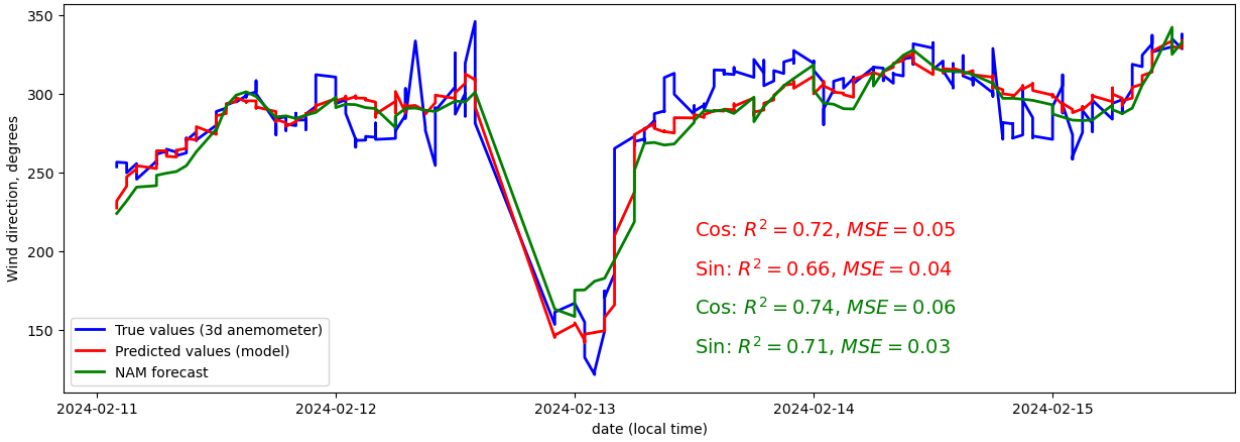

Figure 15. Model values and NAM forecast values and 3D anemometer measurements of wind direction. The model and NAM forecast had roughly the same R^2 and MSE values.

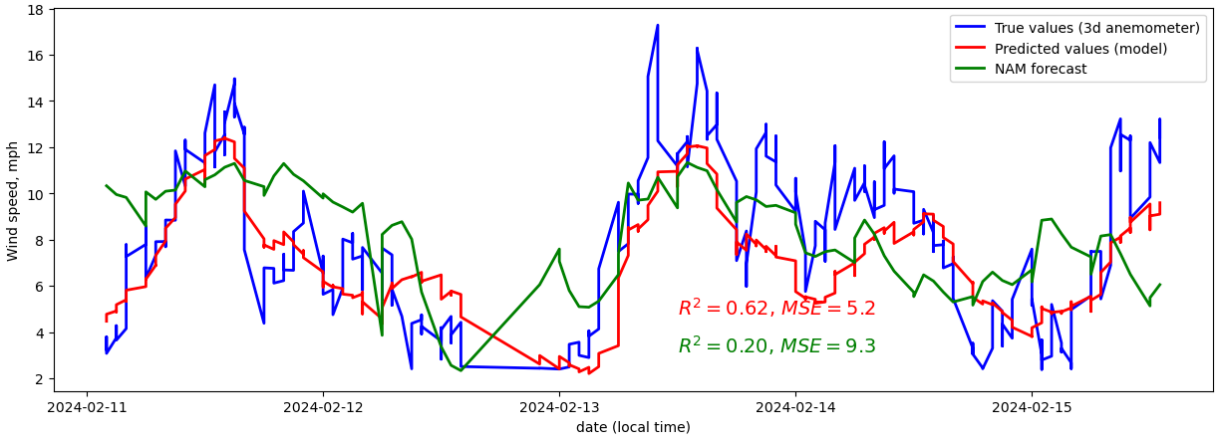

Figure 16. Model values and NAM forecast values and 3D anemometer measurements of wind speed. The model has a higher R^2 value and lower MSE than the NAM forecast.

A note on circular encoding

In setting up the multilinear regression models, it was important to address the circular nature of a few of the variables, namely the wind direction, hour of day, and day of year. Simply using the wind direction as a numeric value (0-359.9), for example, would create an artificial discontinuity at the boundary between 0 and 360. Standard linear regression cannot handle this properly because it assumes a linear relationship and does not account for the circular wrap-around of angles. To address this, the wind direction was broken up into the cosine and sine components. Since 0° represents north, the cosine of the wind direction represents the N-S component (with north being positive), while the sine of the wind direction represents the E-S component (with east being positive). The hour of day was similarly transformed into two components, i.e. $cos\left(\frac{2\pi}{24} * hour\right)$ and $sin\left(\frac{2\pi}{24} * hour\right)$. The day of year would technically need the same treatment, but since the data in this study only spanned 30 days that did not include a change in year, no circular encoding was necessary.

For the predictor variables that required circular encoding, it was easy enough to use the cosine and sine components as predictor variables and leave it at that. However, when a dependent variable requires circular encoding, it becomes a bit more complicated, necessitating two separate multilinear regressions: one for the cosine component and one for the sine component. The predicted value of the desired dependent variable must then be constructed from the predicted cosine and sine components. Since wind direction was a dependent variable in this study, this is exactly what had to be done. The predicted wind direction in degrees was constructed from the predicted cosine and sine components generated from two separate multilinear regression models.

Evaluating the multilinear regression models

To evaluate a model, we can look at the R^2 value and the Mean Squared Error (MSE) between the truth (in this case the 3D anemometer output) and the prediction (in this case, the value predicted by the model). The R^2 value, also known as the coefficient of determination, is a statistical measure used to evaluate the performance of a regression model. The R^2 value represents the proportion of the variance in the dependent

variable that can be explained by the independent variables in the model. A value closer to 1 indicates that the model explains all the variability of the dependent variable. The predictions perfectly match the actual data. The MSE measures the average squared difference between the actual values and the predicted values generated by the model. A lower MSE value indicates that the model's predictions are closer to the actual values, suggesting a better fit. The MSE is expressed in the square of the units of the dependent variable, so it is mainly interpretable when comparing two different models for the same dependent variable. The R^2 and MSE values for the multilinear regression models are shown in *Table 1*.

Table 1. R^2 and MSE values for the multilinear regression models. The higher R^2 and lower MSE values are highlighted in green.

	Temperature, F	Wind direction, cos	Wind direction, sin	Wind speed, mph
R-value model prediction	0.92	0.85	0.81	0.79
R-value NAM forecast	0.82	0.86	0.83	0.44
R^2 model prediction	0.85	0.72	0.66	0.62
R^2 NAM forecast	0.67	0.74	0.71	0.20
MSE model prediction	10.76	0.048	0.036	5.23
MSE NAM forecast	17.25	0.058	0.027	9.29

Let us first look at the multilinear regression model for temperature. This model has a an R^2 value of 0.85, higher than R^2 value of the NAM forecast (0.67). The model also had a lower MSE than the NAM forecast. Both of these results indicate that the multilinear regression model does a much better job at modeling the temperature recorded by the 3D anemometer than the direct NAM forecast for temperature.

Let us next look at the multilinear regression models for wind direction (there were two, corresponding to the cosine and sine components of the wind direction). For the regression on the cosine of wind direction, the model has a an R^2 value of 0.72, slightly lower than the R^2 value of the NAM forecast (0.74). For the regression on the sine of wind direction, the model has an R^2 value of 0.66, slightly lower than the NAM forecast (0.71). The MSE values are comparable. These results indicate that the multilinear regression model does a slightly poorer job at predicting site wind direction than the NAM forecast, but not by much. The R^2 values hovering around 0.7 suggest a fairly good fit, but there is still some variability in the data unexplained by the models and NAM forecast.

Finally, let us first look at the multilinear regression model for wind speed. This model has a an R^2 value of 0.62, significantly higher than R^2 value of the NAM forecast (0.20). The model also had a significantly lower MSE than the NAM forecast. Both of these results indicate that the multilinear regression model does a much better job at modeling the wind speed recorded by the 3D anemometer than the direct NAM forecast of wind speed. However, while the model's R^2 value 0.62 suggests a fairly good fit, there is still some variability in the data unexplained by the model.

Another way to evaluate the multilinear regression models specifically in comparison to the NAM forecasts is to plot histograms of the difference between the model prediction and the actual 3D anemometer measurement and compare these to the histograms of the difference between the NAM forecast and actual 3D anemometer measurement. This is shown in *Figure 17*. For all three dependent variables, the multilinear regression model gives a tighter histogram, with fewer large difference between the predicted and actual

values. This is especially true for the multilinear temperature and wind speed models, which agrees with findings of higher R^2 and lower MSE values for these two models over the NAM forecasts.

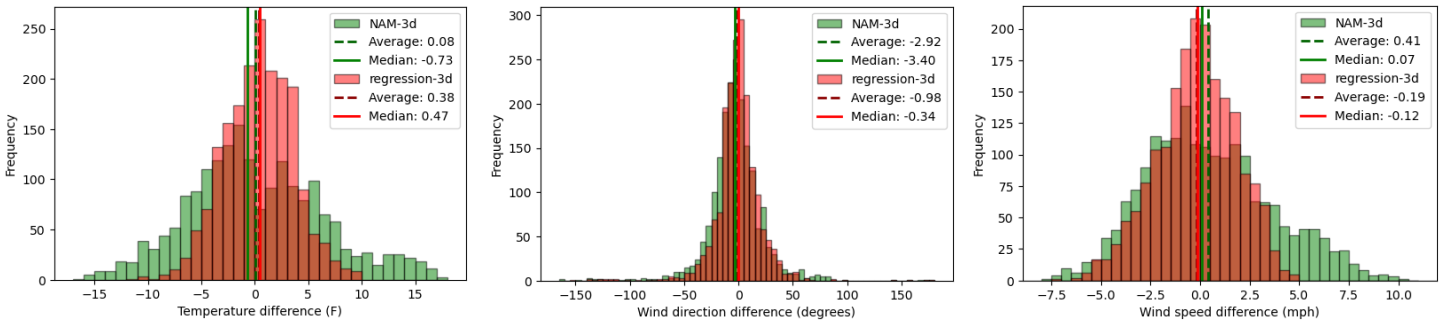

Figure 17. Histograms of the differences between the NAM data and the 3d anemometer and the regression prediction and the 3d anemometer. The forecast period is limited 1-12 hours and the wind speeds > 1m/s (2.24 mph). For all three dependent variables, the multilinear regression model gives a tighter histogram, with fewer large differences between the predicted and actual values. (The NAM-3d is the same data as in Figure 6 histograms for NAM-3d.)

Coefficients of multilinear regression models

It is also instructive to look at the coefficients of the multilinear regression models. Standardized coefficients in a multilinear regression provide a way to compare the relative importance of different independent variables on the dependent variable, regardless of their original scales. The magnitude of a standardized coefficient indicates the strength of the relationship between an independent variable and the dependent variable. Larger absolute values imply a greater impact on the dependent variable. The sign on the coefficient indicates the direction of the relationship. A positive coefficient means that as the independent variable increases, the dependent variable tends to increase; a negative coefficient indicates the opposite. *Table 2* shows the values of the standardized coefficients for each model.

Table 2. Coefficients for standardized variables for multilinear regression models. For a given model, coefficients with a higher magnitude have a greater predictive power on the output. The gradient coloring highlights the coefficients with greater predictive power (darker) in each of the models. Green indicates a positive effect while red indicates a negative effect.

Variable (Standardized)	Temperature, F	Wind direction, cos	Wind direction, sin	Wind speed, mph
temp_F_nam	4.46	-0.17	-0.08	0.19
wdr_cos_nam	-0.89	0.12	-0.01	-0.06
wdr_sin_nam	-0.57	-0.19	0.29	-0.17
wspd_mph_nam	1.33	-0.16	-0.13	0.31
forecast_period	0.04	0.00	-0.03	-0.06
welv_nam	0.00	0.02	0.00	0.04
kinetic_e	-1.13	0.12	0.10	0.36
hour_forecasted_cos	-0.33	0.02	0.04	-0.87
hour_forecasted_sin	0.75	0.03	0.00	-0.48
day_of_year_forecasted	-0.63	0.03	-0.03	0.07
temp_F_nam_lag_b1	2.24	0.00	0.01	-0.03

temp_F_nam_lag_b2	1.20	0.00	0.06	-0.17
temp_F_nam_lag_f1	2.30	-0.04	0.02	-0.09
temp_F_nam_lag_f2	2.91	0.16	0.00	-0.39
wdr_cos_nam_lag_b1	-0.46	0.06	0.04	0.22
wdr_cos_nam_lag_b2	-0.03	0.17	-0.03	1.07
wdr_cos_nam_lag_f1	-0.40	0.05	-0.05	-0.18
wdr_cos_nam_lag_f2	-0.35	-0.02	-0.03	-0.41
wdr_sin_nam_lag_b1	0.20	-0.06	0.23	0.39
wdr_sin_nam_lag_b2	0.68	0.08	0.15	1.03
wdr_sin_nam_lag_f1	-0.54	0.04	0.02	-0.11
wdr_sin_nam_lag_f2	-1.04	0.02	-0.07	-0.62
wspd_mph_nam_lag_b1	1.06	0.04	-0.02	-0.01
wspd_mph_nam_lag_b2	0.79	0.07	-0.05	0.24
wspd_mph_nam_lag_f1	1.27	-0.09	0.00	0.56
wspd_mph_nam_lag_f2	1.41	0.09	0.07	0.01
kinetic_e_lag_b1	-1.04	-0.04	-0.01	0.57
kinetic_e_lag_b2	-0.46	-0.04	-0.04	0.20
kinetic_e_lag_f1	-1.45	0.08	0.01	0.25
kinetic_e_lag_f2	-1.62	-0.06	-0.01	0.62

Let us first look at the multilinear regression model for temperature. The strongest predictors of temperature are the NAM forecast temperature (no surprise), kinetic energy, and wind speed. The corresponding lag variables for both before and after the time in question also have predictive power. Looking at the signs of the coefficients, we see that increasing NAM forecast temperatures, increasing NAM forecast wind speeds, and decreasing NAM forecast kinetic energy will all lead to an increase in the predicted temperature at the location of the 3D anemometer.

Let us next look at the multilinear regression models for wind direction (there were two, corresponding to the cosine and sine components of the wind direction). The strongest predictors of wind direction are the NAM forecast components of wind direction (no surprise), wind speed, and temperature. Because of the two components, it is a bit less intuitive to see how wind direction itself changes with increasing or decreases in the independent variables.

Finally, let us first look at the multilinear regression model for wind speed. The strongest predictors of wind speed are the NAM forecasted wind direction components before the time in question, the kinetic energy just before the time in question, and the hour of day. Interestingly, the actual NAM forecast wind speed is not a super strong predictor for the wind speed recorded by the 3D anemometer. Looking at the signs of the coefficients, increasing NAM forecast kinetic energy, increasing the cosine or sine component of the wind, or decreasing the cosine of the hour (the hours with the smaller cosines are noon and the hours surrounding it) will all lead to an increase in the predicted wind speed at the location of the 3D anemometer.

Conclusions

Is forecast data a good proxy for anemometer data?

Based on this analysis, the NAM forecast data for temperature, wind direction, and wind speed has a moderately strong correlation with the anemometer measurements. There is enough scatter that the NAM data is likely not viable as a direct replacement of the anemometer data on a 15-minute time frame. The NAM forecast could be used to predict such things as plume movement, for example, over time.

Does a multilinear regression trained on forecast data provide an even better proxy for anemometer data?

While the direct NAM forecast may not be a sufficient consistent proxy for anemometer data, predictions based off a site-specific trained multilinear regression model of the publicly-provided NAM forecast data might be, at least during times when the anemometer is inoperable and when the wind speeds are high enough. In this study, multilinear regression models were developed to predict temperature, wind direction, and wind speed at the site of the 3D anemometer. The predicted temperatures and wind speeds offered significantly better correlation and fewer large differences with the actual measurements than the direct NAM forecasts of temperature and wind speed, while the predicted wind directions were about the same. Both the direct NAM forecasts and the models offer the best predictions when the wind speed is above 1m/s. However, there is still enough variation in the site-specific conditions that is not captured by the models that the model predictions are not good enough to just replace the use of anemometers altogether.

Further Study

Where to go next?

This analysis is based on a 30-day span of measurements at a single location in North Dakota. Due to wintery conditions leading to a frozen anemometer, there were 26 days for the 3D anemometer over this 30-day window. The multilinear regression models were developed on a dataset of just 1912 points. Moreover, the measurements were taken in February and March, when weather can be quite harsh, and might lead to less accurate measurements or more difficult forecasting. So, a natural next step for this study would be to extend the 3D anemometer dataset for a longer period of time, ideally throughout an entire year, and rerun the analyses. Also, data from other locations would be equally useful. The multilinear regression model has huge potential to provide good forecasting on a site-specific basis, and it could be made more robust with more data on which to train.

References

3D anemometer data. Courtesy of LongPath Technologies, Inc. Boulder, CO. https://www.longpathtech.com/

North American Mesoscale Forecast System (NAM) data on a 12km grid for all of North America. https://www.ncei.noaa.gov/products/weather-climate-models/north-american-mesoscale

Wind Profile Calculator. The Swiss Wind Power Data Website. https://wind-data.ch/tools/profile.php

Figure of NAM 12km grid. *Enhancing Weather-Related Power Outage Prediction by Event Severity Classification.* By Yang, Watson, Koukoula, Anagnostou. March 2020. https://www.researchgate.net/figure/NAM-and-WRF-domains_fig1_340181673

Air Pressure at Altitude Calculator. https://www.mide.com/air-pressure-at-altitude-calculator