Nonlinear Statistical Models for Predicting Stations Surface Fields

Xinyi Wang, Yang Cheng, Jim Ritter

Sponsor representatives: Razvan Stefanescu, Dusanka Zupanski

02/25/2019-05/14/2019

1 Abstract

The Spire Nonlinear Statistical Models for Predicting Stations Surface Fields Project evaluated the use of a neural network model to predict surface wind speeds. The team used National Weather Service (NWS) Global Forecaast System (GFS) weather forecast data, NWS observed weather measurements from weather stations in Colorado, and SKLearn neural network modules to determine if it was possible to predict wind speeds at the surface. The team gathered multiple predictors from the NWS forecast data and gathered observed wind speeds as predictands to use to train and test the neural network model. The neural network model was able predict wind speeds effectively when the optimal set of predictors were processed.
[1] [2] [3]

2 Introduction

The Spire Nonlinear Statistical Models for Predicting Stations Surface Fields project was set up to evaluate the possibility of using a neural network (nonlinear) model to predict surface wind speeds. The objective of the project is to use a Neural Network to predict surface wind speeds with a better accuracy than the forecast wind speeds from the NWS. The method is similar with Mao and Monahan, and Mean Abosolute Error(MAE) and Root mean squared error(RMSE) are going to be used to verify the model(Mao, Monahan).

Surface winds are a climatic field of interest since they play an important role in many industries including agriculture, transportation and new energy. Since each weather forecast is related to four weather stations, if we based on the

[1] Xinyi Wang - Analysis with linear regression and reporting
[2] Yang Cheng - Neural Network set up, adjustments and reporting
[3] Jim Ritter - Data input, correlation and reporting

neural network to find best relationship between them and predict as accurate as possible, it will help Spire and other wind related stakeholders better prepare and manage their company.
The team paired data from Global Forecast System(GFS) geometric grid points with observed data from weather station locations in Colorado. The team would use results from a neural network to compare with a linear model, and observe the differences between weather forecasts and real station observations.

3 Methods

3.1 Overview of the Method

The method that the team employed was to use a neural network to predict surface wind speeds. The first step was to gather forecast data from files from the GFS system to use as predictors and , and then to gather observation data to use as predictands for training. The team created a dataset with forecast variables that would be used as predictors in the neural network, and as input for linear regression models. The team would do analysis with the forecast data using linear regression models for comparison to the non-linear method.

The project team received direction from the Spire team for the objective of the project, and for a high level approach to use. The Spire team provided links to the NOAA site to get forecast files from the National Weather Service(GFS). They also provided weather observation data taken from weather stations across Colorado. The Spire team provided explanation for the file format of the files from NOAA, along with a python module to use to read files. The Spire team suggested that the project team should gather forecast data, and observation data, and work out a method to compare the two.

3.2 Describe the data

The team would not be able to directly compare surface wind speeds from the forecast files to the wind speeds in the observation files because the locations for the two did not match. The Forecast data is created for locations on grid points that are laid out with Latitude and Longitude lines. The observation data is gathered at weather stations that are not aligned with teh forecast grid. The elevation for the observed wind speeds is the surface elevation of the weather station in meters above sea level. The elevation for forecast wind speeds is 10 meters above the surface, at each forecast grid point.

The Forecast Grid is spaced by one half longitude and latitude lines. For example, there is a point at 37 degrees north latitude, by 255 degrees west longitude, and another at 37.5 degrees north by 255 degrees west.

Forecasts are created 4 times a day, with varying forecast lead times. The

observation files contained observations from multiple times during the day, but most were between 10:00 am and 1:00 pm. For the purposes of this project, the team elected to use observation data taken at 12:00 each day, and forecast data that was produced at 6:00 am everyday, with a 6 hour forecast time, so that the forecast is for 12:00.

The weather observations for this project are taken at various locations across Colorado. The locations, as a rule, do not line up with the Grid Points from the forecast files. The weather stations that provide weather observation data in Colorado are typically located near larger cities and towns, and in a lot of cases are located at an airport. They are not evenly distributed across the state.

3.3 Clean up and organize the Data

The project team wrote code to correlate wind speeds between the station locations, and the forecast grid points using 3 different methods.

1. The first method was to start with the weather station location and calculate the nearest forecast gridpoint to the location. The observed wind speed from the weather station would be compared to the forecast wind speed at the nearest forecast grid point.

2. The second method was to find the four grid points surrounding the weather station location, calculate the average of the wind speeds forecast for those 4 grid points, and compare the observed wind speed from the weather station.

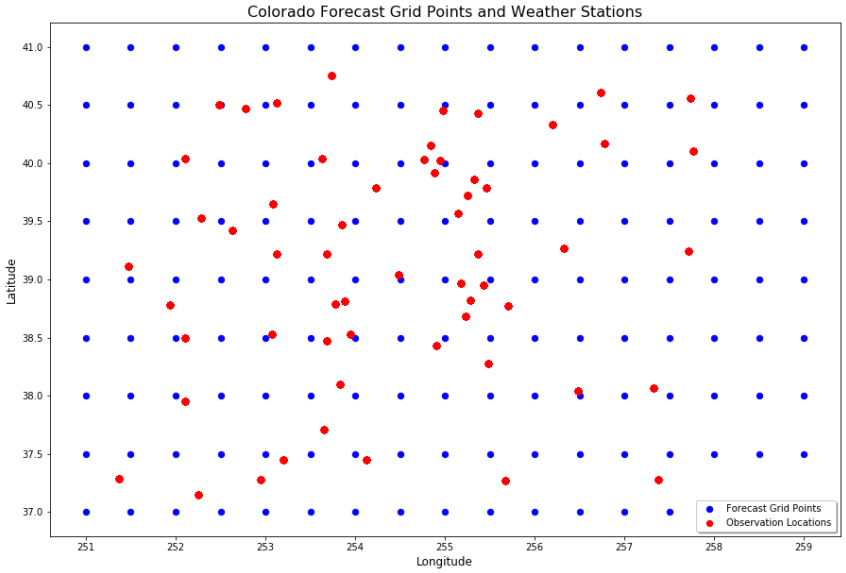
3. The third method was to find the four grid points surrounding the weather station and calculate a weighted average of the wind speeds forecast for those 4 grid points. The weight is determined by using a bilinear interpolation formula. The weighting provides higher influence for grid points closer to the weather stations. The weighting that was used was :

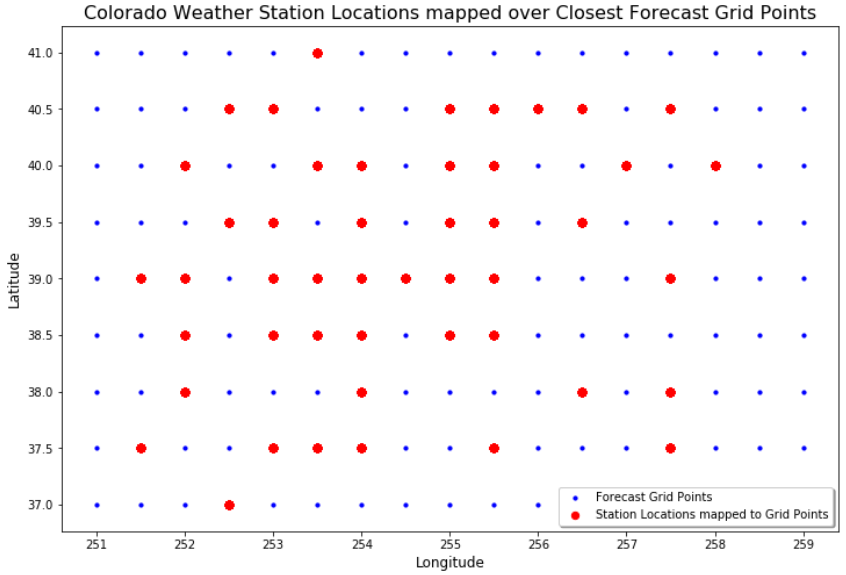
$$\begin{aligned} Weightedavg = \tfrac{1}{(x_1-x_2)(y_1-y_2)} \times f(x_{11}) \times ((x_2-x) \times (y_2-y)) + \\ f(x_{21}) \times ((x-x_1) \times (y_2-y)) + \\ f(x_{12}) \times ((x_2-x) \times (y-y_1)) + \\ f(x_{22}) \times ((x-x_1) \times (y-y_1)) \end{aligned} \tag{1}$$

   (a) Where x is the longitude of the weather station, y is the latitude of the weather station, x_1 is the longitude of the grid points to the left (west) of the weather station. x_2 is the longitude of the grid points to the right (east) of the weather station. y_1 and y_2 are the latitudes of the grid points above and below (north and south) respectively. $f(x_{11})$ is the forecast wind speed at the grid point (x_1, y_1), $f(x_{12})$ is the forecast wind speed at the grid point (x_1, y_2) , $f(x_{21})$ is the forecast wind speed at the grid point (x_2, y_1) , and $f(x_{22})$ is the forecast wind speed at the grid point (x_2, y_2). The wind speeds are all measured in meters per second.

.

The following figures illustrate the location of weather stations in Colorado, and their proximity to forecast grid points, and the mapping of weather station locations to the nearest forecast grid points.

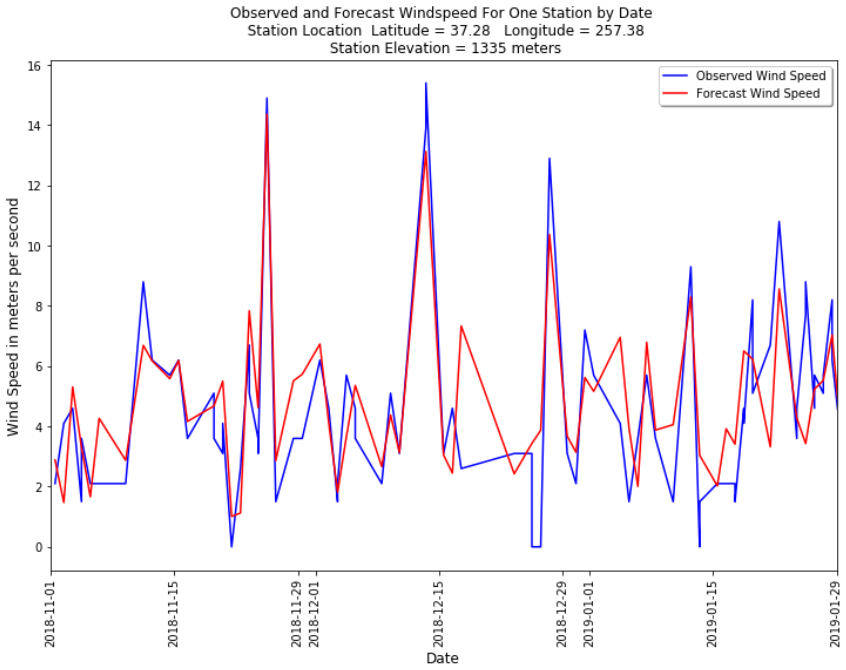




3.4 Analysis

After cleaning up duplicate entries, and merging weather forecast data and observation data, the team analyzed the merged dataset and created a linear model.

1. Firstly, the team compared forecast and observed wind speeds by graphing both wind speeds over time for each weather station.



2. Then, the team used five number summary to detect min, first quartile, median, third quartile and max for forecast wind speed and observation wind speed.

3. The team extracted two different wind speeds separately from the dataset and generated a box plot for them.

4. Then the team generated a scatter plot to show whether it is linear or non-linear relationship between observation wind speed and weather forecast wind speed.

5. The relationship between observation wind speed and weather forecast wind speed, is not obvious because of the small distance. Therefore, instead of drawing scatter plot, the team tested test the relationship with a linear regression model.Here the team used the merged dataset.

$$WS_{obs} = 1.59 + 0.58WS_{for} + \epsilon \tag{2}$$

(a) WS_{obs} represents observation wind speed, WS_{for} is weather forecast wind speed. Adjusted R^2=0.517 (The adjusted r-square is a standardized indicator of r-square, adjusting for the number of predictor variables, the value is between 0 and 1.) 0.517 is not good enough, then the team tried backward method.

3.5 Backward Method

1. Firstly, the team generate a full model, which includes not only WSfor but also Temp, PRMSL,and PRES.

$$WS_{obs} = 43.45+3.75\times10^{-3}Elevation+0.75WS_{for}-5.53\times10^{-2}Temp-5.62\times10^{-4}PRMSL-2.80\times10^{-4}P \quad (3)$$

   (a) AdjustedR^2=0.686 It is bigger than the first regression, there is a better model.

2. Next it is necessary to delete each variable. After several comparison, it is the most efficient to delete PRMSL.

$$WS_{obs} = -27.99+3.68\times10^{-3}Elevation+7.60WS_{for}+9.72\times10^{-3}Temp-2.39\times10^{-4}PRES+\epsilon \quad (4)$$

   (a) Adjusted R^2=0.671 It is a little bit smaller than the full model, but condition number (Measures how much the output value of the function can change for a small change in the input argument) is much smaller.

3.6 Identifying Unusual Observation

After backward method, now it is necessary to solve other numerical problem by identifying unusual observation.

1. Find out influential observation(which is one that causes a substantial change in the fitted model based on its inclusion or deletion from the model).

$$WS_{obs} = -27.21+3.88\times10^{-3}Elevation+0.76WS_{for}+4.46\times10^{-5}Temp+2.57\times10^{-4}PRES+\epsilon \quad (5)$$

3.7 General explanation of Neural Network method

The SKLearn neural network module requires numpy arrays for it's input. It requires an array of predictors, and an array of predictands to use for training and testing. The length of the two arrays must be the same. The predictor arrays will have several entries. In this project the predictor arrays have forecast wind speed, forecast temperature, forecast pressure reduced to mean sea level, and forecast pressure, the predictand array only has observed wind speeds.

For the Neural Network, we feed input data into the neural network, then the information is delivered between the sections from one layer to another, finally outputting the prediction. The Neural network can be divided into supervised learning and unsupervised learning. Supervised learning means learning from paired input-output datasets. Feed Forward Neural Network is the model we used for machine learning. Considering the task is predicting the wind speed, supervised learning is the one we want to choose. To train the model, comparison between the prediction with the real data is necessary. A loss function is used to measure the difference and is the primary criterion to train the model. The loss function to be used in model below is MSE-Mean Square Error. MSE provides a good property to compare predictivity with linear model. However, when used with sigmoid function, MSE will decrease the learning speed. To train the model, Gradient Descent and Backward propagation is needed. Gradient descent is to decrease a small interval times the derivative of that point to change the weight find the local minimization of loss function. Backward propagation algorithm is used to distribute error in output to previous layer in order to adjust different weight.

Neural Network is one of the non-linear models, which means it has better ability to extract relationship when response variable is not the linear combination of its explanatory variable. To introduce non-linearity to the model, neurons need to be processed by activation function. The commonly used activation function is sigmoid function and tanh function.

3.8 Feed Forward Neural Network

There are three components of neural networks including input layer, hidden layers and output layers. The number of nodes in input layer is decided by the number of predictors. The output layer only contains one node which is the wind speed. There is no best way to choose the number of hidden layers and number of nodes in hidden layer. Hence in our experiment, they are decided by rules of thumb, including number of hidden nodes is no more than twice of the number of nodes in input layer. Also, more than 2 hidden layers is not necessary. After that we just test performance of different nodes and hidden layer by R square and correlation between prediction value and real value to decide the final structure.
The Neural Network needs to be trained to adjust the weight between different layers. The number of trained the model experiences is called epoch. However, too many epochs will cause the model loss generality, which is called overfitting problem. Early stopping is used to avoid overfitting problem. By dividing data into two validation set and training set, we can compare the performance of those models which is trained training set and test validation set. By drawing a graph with x to be epoch and y to be the loss function, we can find the minimum validation lost in the model. The generality will be lost when adding more epochs. The early stopping method takes too much computation power

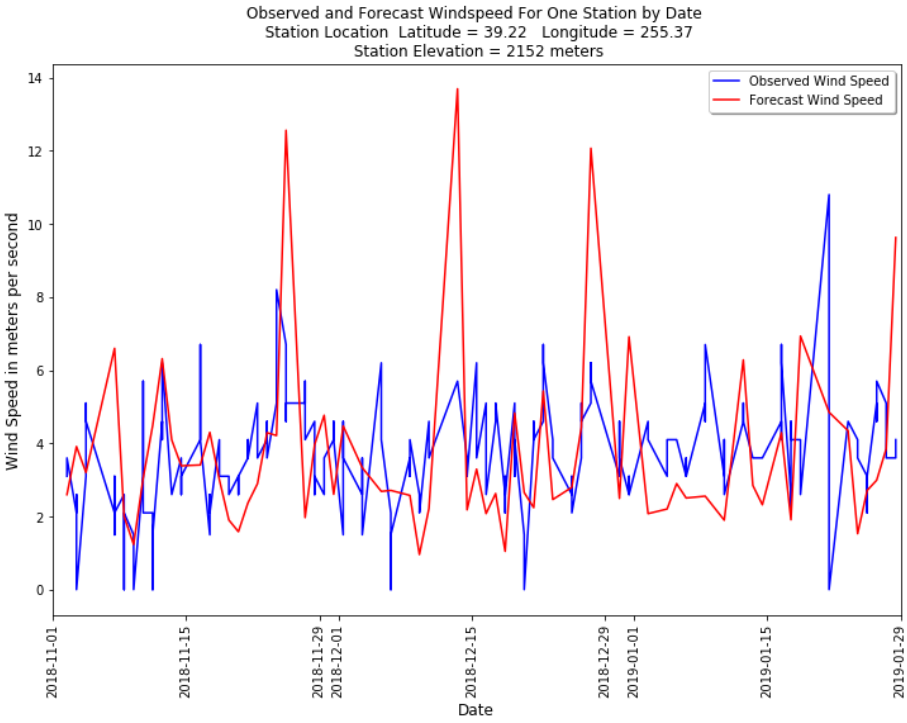
and we failed to test it with more than 200 epochs. However, 200 epochs are far from the minimum validation. This is not the problem because the fitting of mode in validation data is usually less than fitting of model in training data.

4 Results

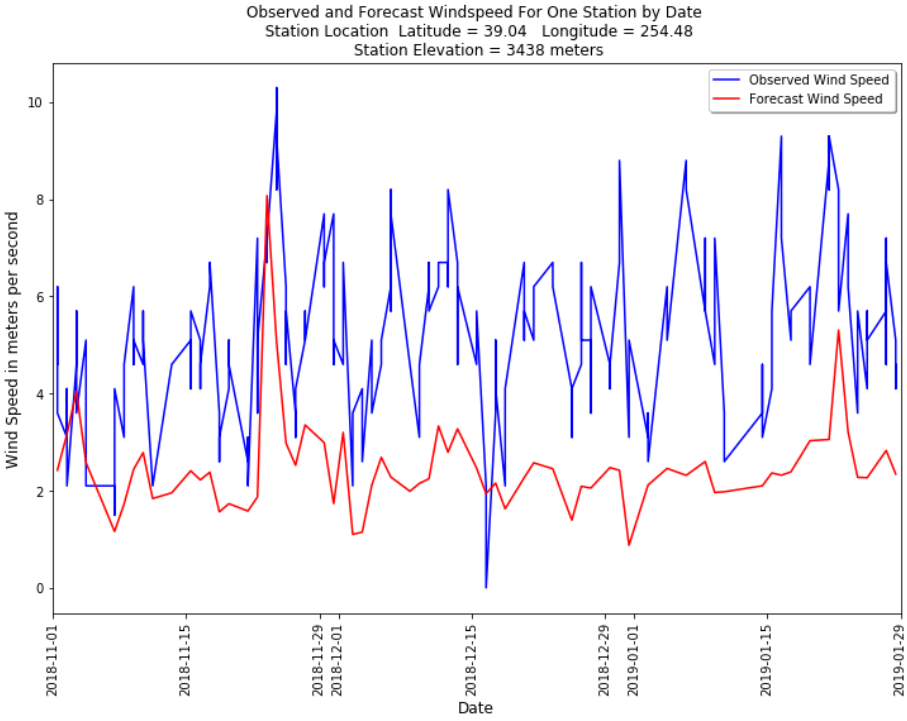
The Jupyter Notebook that the team wrote creates a file with observed wind speeds, as well as the above mentioned varieties (closest grid point, average, and weighted average calculations) of forecast wind speed, temperature, pressure reduced to mean sea level, pressure, and elevation. All three sets of forecast values are extracted for use as predictors in the neural network steps. The neural network was able to predict wind speeds with improved accuracy when elevation was added as a predictor.

The graphs of forecast wind speeds vs. observed wind speeds showed good correlations for most weather stations at lower altitudes. For the weather stations at higher altitudes, especially weather stations above 3000 meters, the observed wind speeds were typically much higher than the forecast wind speeds. There also was a situation where some weather stations displayed "suspect" wind speeds of zero for prolonged periods that appear abnormal. These two situations set up a scenario where the difference between observed and forecast wind speeds are inflated.

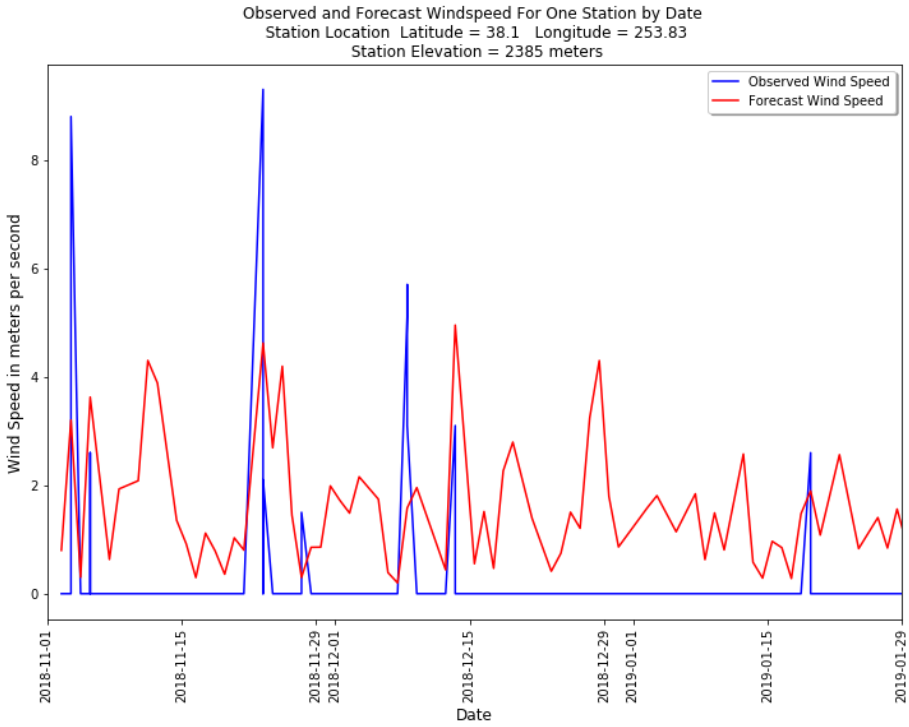
This weather station at 2152 meters shows a fairly close correlation between observed and forecast wind speeds.



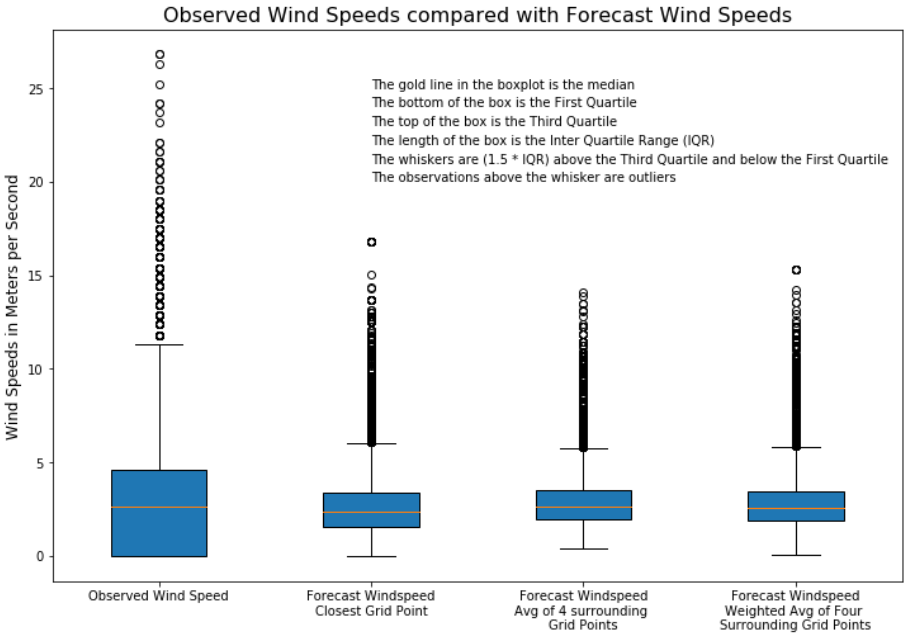
This weather station at 3430 meters shows a pattern where the observed wind speeds are significantly higher than forecast wind speeds.



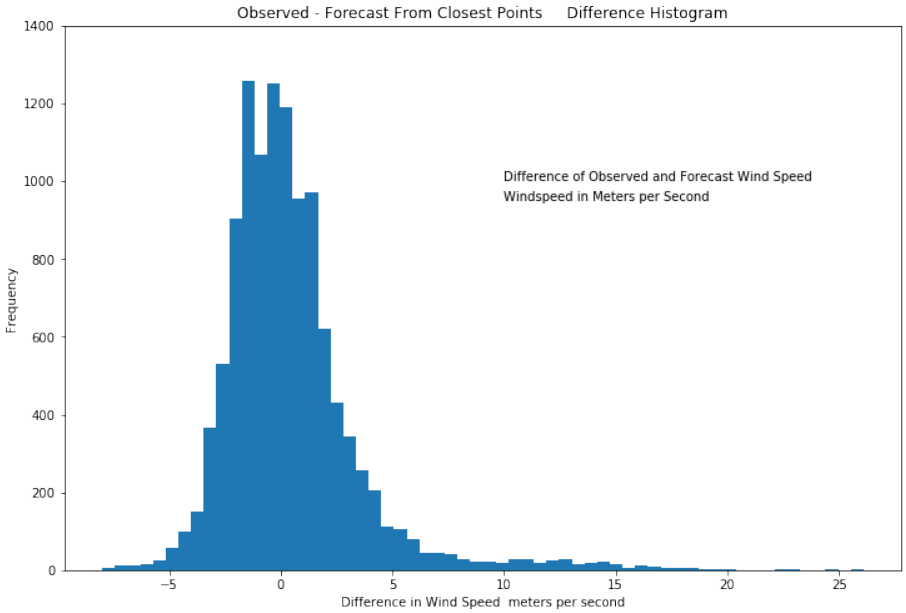
This weather station seems to have an abnormal number of wind speed observations at zero.

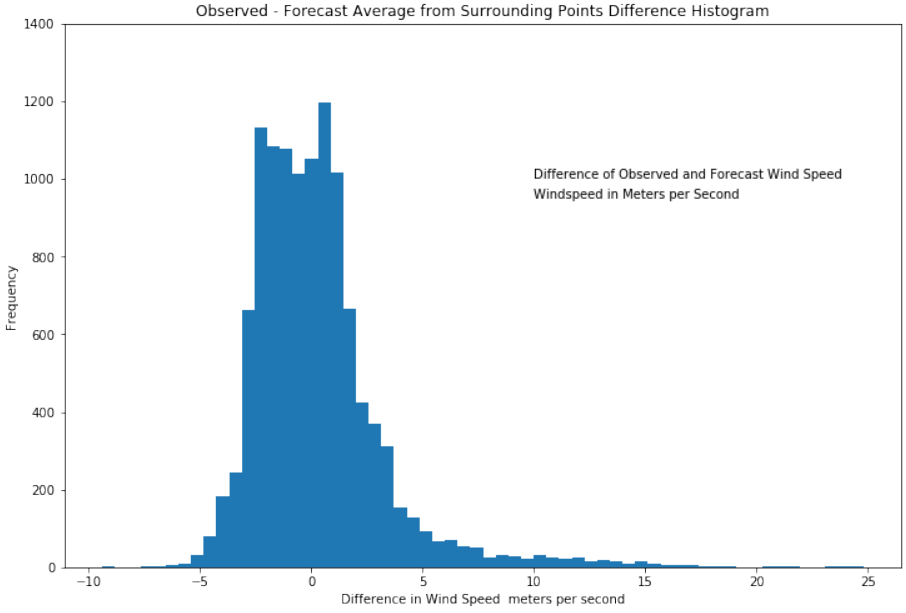


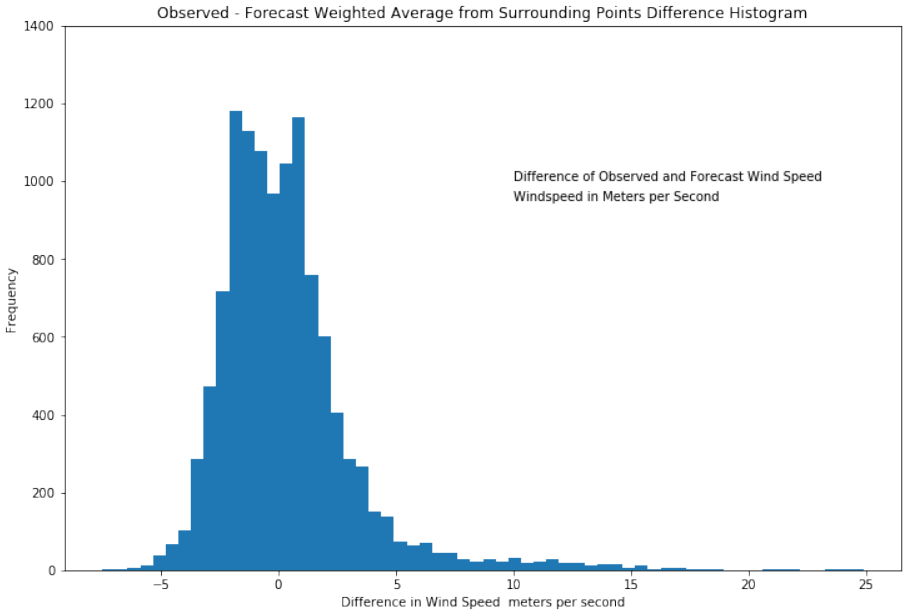
The box plots of the observed wind speeds shows a larger IQR in comparison to the boxplots of forecasted wind speeds.



Histograms of the differences of Observed Wind Speeds and Forecast Wind Speeds are centered on zero, and are distributed fairly normally, but are skewed to the right indicating some higher Observed Wind Speeds.


Observed - Forecast From Closest Points Difference Histogram
Difference of Observed and Forecast Wind Speed
Windspeed in Meters per Second
Frequency
Difference in Wind Speed meters per second
0
200
400
600
800
1000
1200
1400
−5
0
5
10
15
20
25


Observed - Forecast Average from Surrounding Points Difference Histogram
Difference of Observed and Forecast Wind Speed
Windspeed in Meters per Second
Frequency
Difference in Wind Speed meters per second
0
200
400
600
800
1000
1200
1400
−10
−5
0
5
10
15
20
25


Observed - Forecast Weighted Average from Surrounding Points Difference Histogram
Difference of Observed and Forecast Wind Speed
Windspeed in Meters per Second
Frequency
Difference in Wind Speed meters per second
0
200
400
600
800
1000
1200
1400
−5
0
5
10
15
20
25

For weather forecast wind speed:

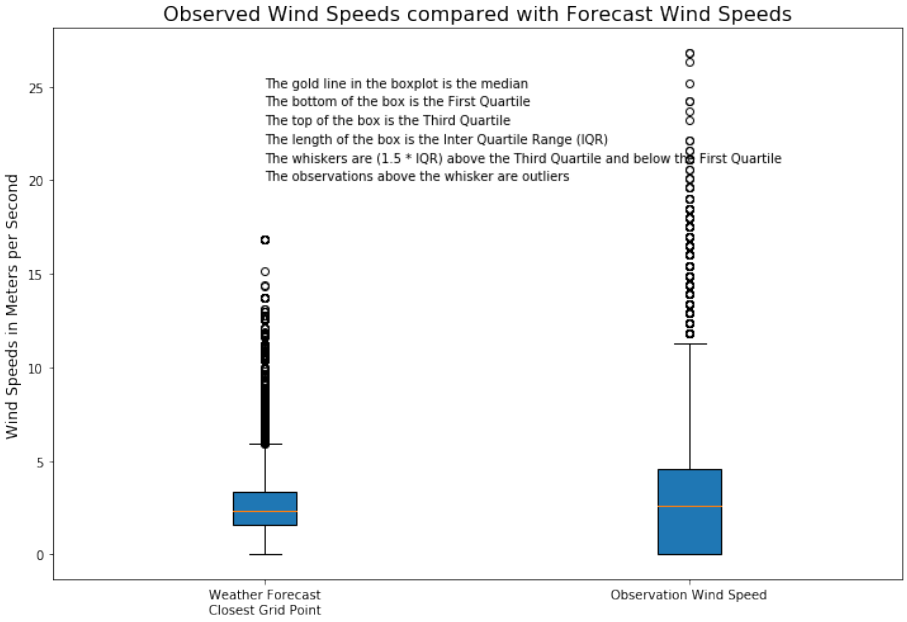
1.

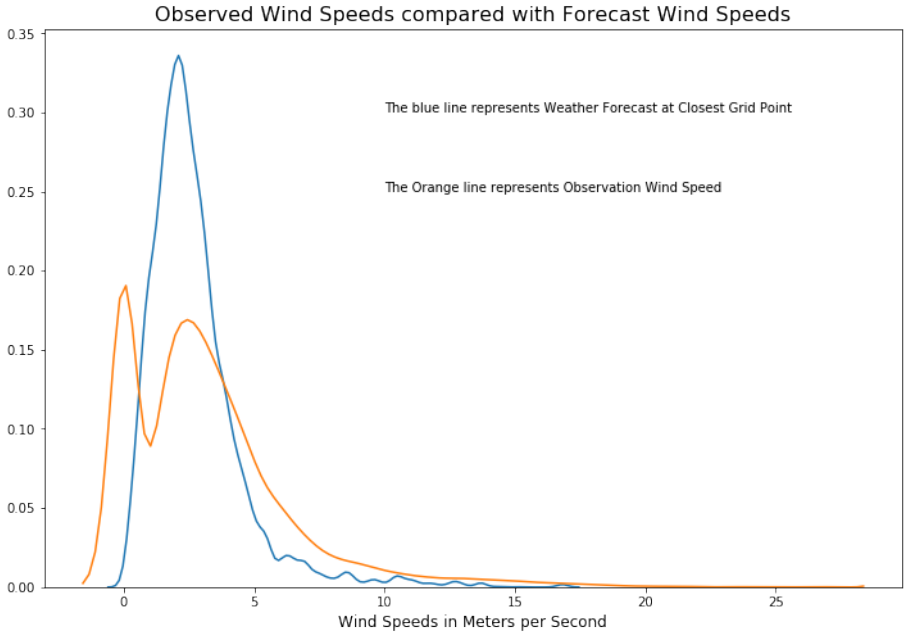
Min	0.020	m/s
Q_1	1.657	m/s
Median	2.418	m/s
Q_3	3.417	m/s
Max	14.360	m/s

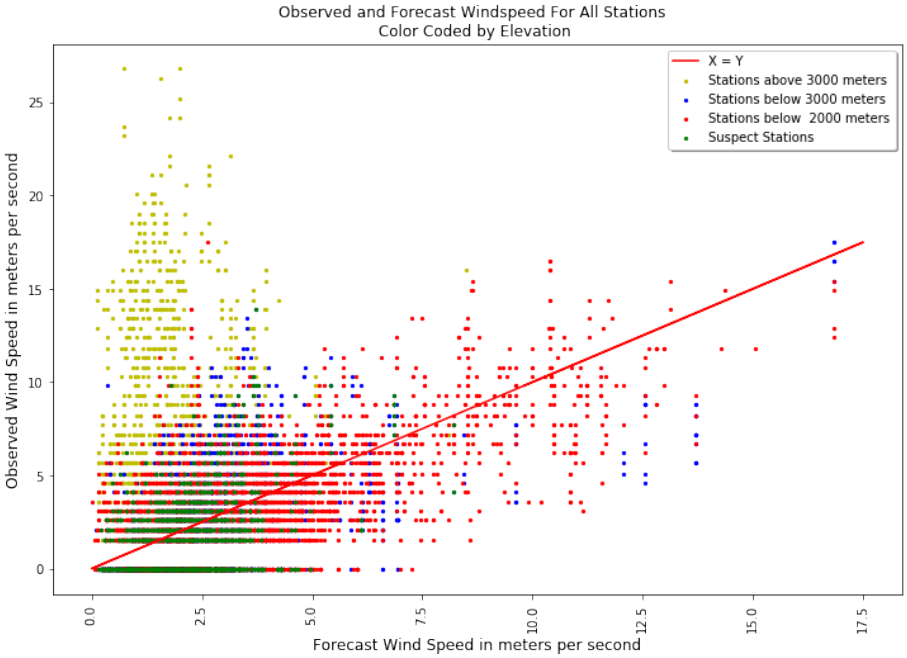
For observation wind speed

1.

Min	0.000	m/s
Q_1	1.500	m/s
Median	2.600	m/s
Q_3	4.600	m/s
Max	26.800	m/s







1. Parts of Elevation which are greater than 3000m are influential, and the

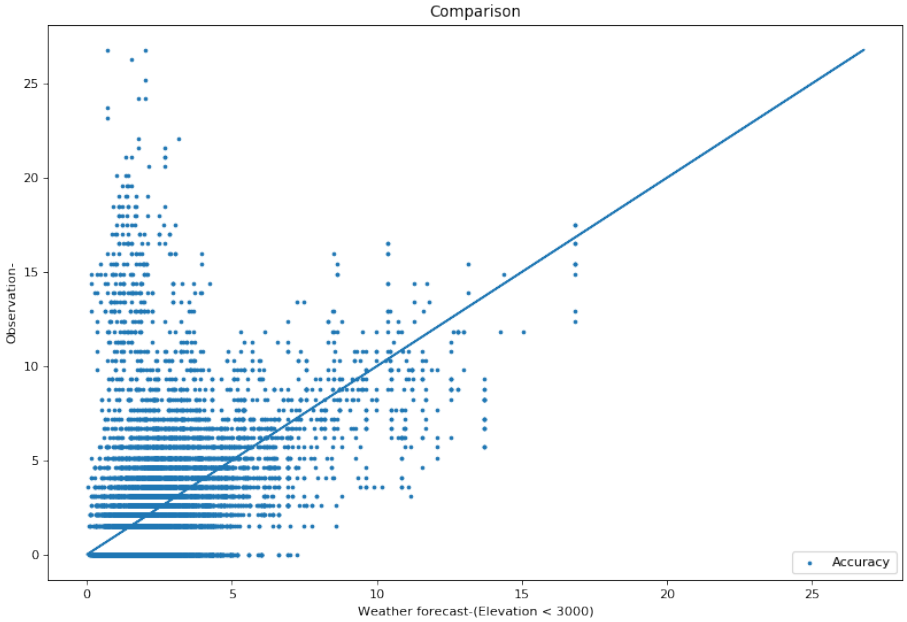
team choose to delete them.

(a) Deleting some of the index will lower down the value of adjusted R^2,
we will only delete index between [720:1007],[2465:2660],[4799:5072],[5306:5415]and[6802:7065]

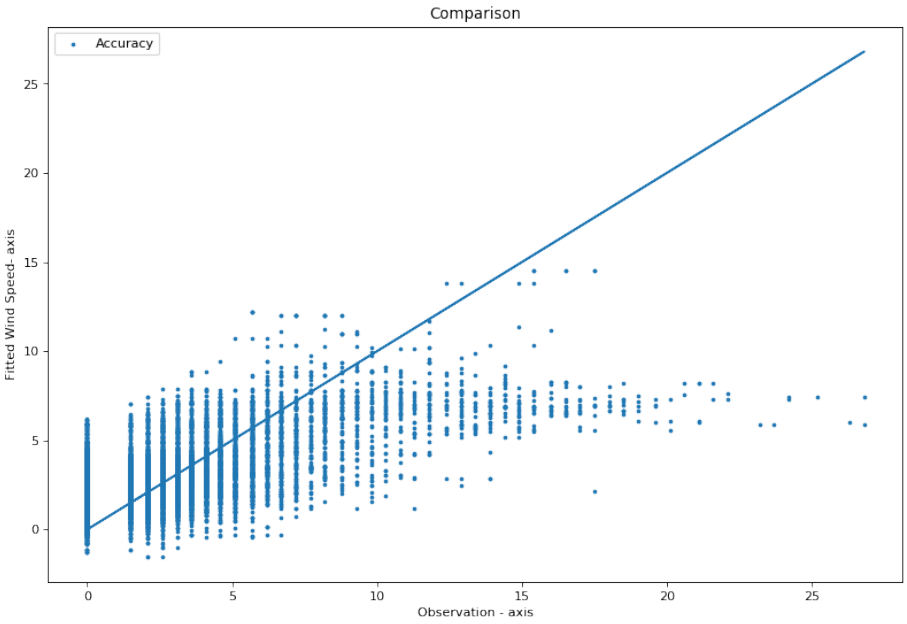
1. At first, the coefficient of full model.

β_0	43.45
β_1	3.75×10^{-3}
β_2	0.75
β_3	-5.53×10^{-2}
β_4	-5.62×10^{-4}
β_5	-2.80×10^{-4}

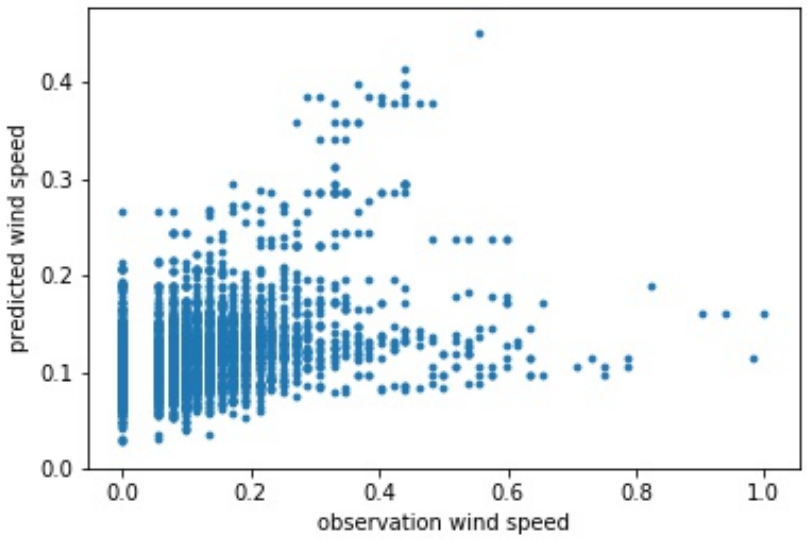
After deleting the index of Elevation > 3000



The observation and fitted model (5)

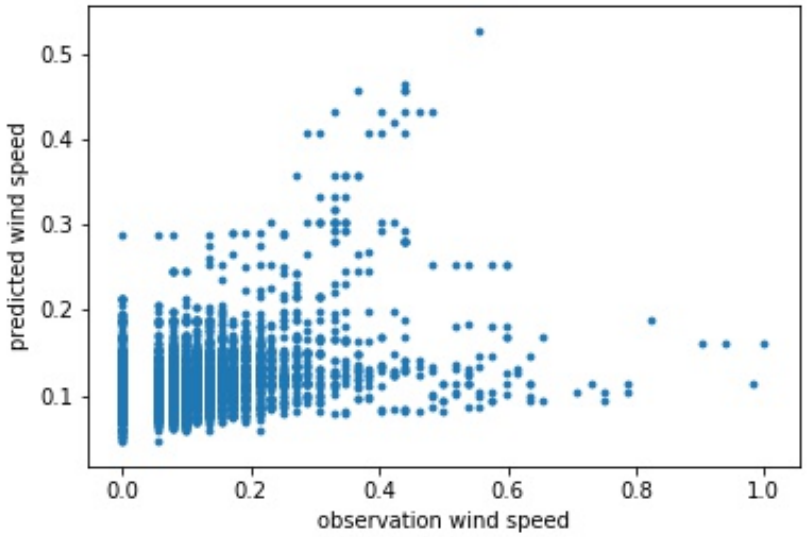


Comparing the neural network model with structure 4-3-1, 4-4-1, 4-5-1, 4-6-1, 4-7-1, 4-8-1. We find out that model with structure 4-5-1 has better performance than other which have r-square to be 0.12. and correlation to be 0.35.



By adding another layer and test 4-5-1-1, 4-5-2-1, 4-5-3-1, 4-5-4-1, 4-5-5-1. The

model with structure 4-5-3-1 is proved to have worse performance with r-squarer to be 0.09 and better performance with correlation to be 0.39.



Need to mention that running the same training different time will get different r-square because the initial weight and bias is random and with gradient decent method they may fall into different local minimum.
Different activation function including tanh and sigmoid is tested. However, the output of sigmoid function is bad because the property of sigmoid function is conflict with the property of mean square error.

5 Discussion

The team compared results from linear regression and the neural network, the neural network was able to produce better results. The wide disparity of wind speeds from the stations at higher elevations, along with the stations that reported abnormal zero wind speeds created a scenario that made correlation of the wind speeds difficult, and affected the ability to do linear regression. That disparity affected early attempts with the neural network to predict wind speeds. When the team added the elevation as one of the predictor variables, the neural network was able to improve on it's performance.

1. Five number summaries for weather forecast wind speed and observation wind speed. The maximum of observation wind speed is far more than the maximum of weather forecast wind speed.

2. The whiskers of box plot are $Q_1 - 1.5IQR$ and $Q_3 + 1.5IQR$, the data beyond those two line are considered as outliers.

3. The scatter plot can saturate with overlapping markers that make density obscure. Therefore, density plot can generate the relationship between observation wind speed and weather forecast wind speed.

   (a) The shape of weather forecast wind speed also shows a problem (Elevation).

4. The scatter plot with line set observation wind speed equals weather forecast wind speed.

5. Using four different colors to represent different stations' elevation. Those with elevations greater than 3000m typically cause a substantial change in the fitted model.

6. After deleting the certain index of elevation, the new scatter plot and line represents observation wind speed and weather forecast speed seems fitted better than before. Adjusted R^2=0.675 makes sure that model improvement.

7. Finally, the scatter plot shows relationship between observation wind speed and model (5) fitted value.

8. For complicated data, no simple transformation or basic linear regression may capture the relationship between the response and regressors. We will skip transformation.

9. We have no absolute confidence to delete observation=0 because there is a chance that wind speed iz zero for a particular day. We cannot transform WSobs into log(WSobs), even though the pattern suggests a log relationship between WSobs and WSfor.

6 Conclusion

The team learned how to use Python to read in, clean up and analyze data. The team also learned a lot about using a neural network to predict temperature and wind speed from others' articles. The team merged two datasets together (Observation and Forecast), analyzed the correlation of the two, and determined the best method to forecast wind, which is closest station wind speed.
The team generated a linear regression model for weather between forecast wind speed and observation wind speed. After using a backward method, transformation, and comparing R2 , the best model was the linear regression model. To better predict the weather by neural network, elevation was added to the predictors, and the team created a larger dataset with more observations and forecast data points, and used more time for model training. A time series model such as ARMA model and recurrent neural network can be used to test

the performance.

7 Required Files

Forecast Files - one for each day to be processed
https://nomads.ncdc.noaa.gov/data/gfs4/201901/20190101/$gfs_{42}0190101_{0}600_{0}06.grb2$

$Observation files - one for each day to be processed$
$/spire/Input_{D}ata/obs$20181101.$xlsx$

$GridPointfile/spire/Input_{D}ata/Grid_{L}ocations.xlsx$

8 Bibliography

Alonzo , Plougonven , Mougeot , Fischer , Dupre and Drobinski ,"From Numerical Weather Prediction outputs to accurate local surface wind speed : statistical modeling and forecasts.", FORWER2017 011, v1, November 8 2017

Yiwen Mao, Adam Monohan,"Linear and nonlinear regression prediction of surface wind components", Climate Dynamics (2018) 51:3291–3309,https://doi.org/10.1007/s00382-018-4079-5, January 2018.

Caren Marzban,"Neural Networks for Postprocessing Model Output: ARPS", July 2002, final form September 2002.

Oliver Fuhrer,Fred Castruccio, Karin Meier-Fleischer, Juerg Schmidli, Jeff Whitaker, Louis Wicker,"Downloading and installing PyNGL and/or PyNIO",https://www.pyngl.ucar.edu, NCAR, date unknown.

"Global Forecast System (GFS), https:/www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs, NOAA National Centers for Environmental Information, date unknown.