Statistical Comparison of Wind Forecast Models

Andrew Sheerin (100672990)

ISE 533 Final Project

University of Rhode Island

5/1/2024

Table of Contents

Table of Tables

Table of Figures

1 Introduction

1.1 Background

Accurate wind forecast models are of significant value to many industries, including wind and renewable energy, athletics, agriculture, transportation, or simply for everyday life. These forecasts provide critical information to aid decision-making that could have important economic, environmental, or social impacts. Personal motivation stems from my passion for sailing, a sport that relies solely on the wind, making reliable wind forecasts an asset. There are multiple wind forecast models that are used globally; however, the forecasted wind speed and direction are sometimes inaccurate when compared to observed data. Moreover, each wind model may present discrepancies between each other, making it difficult to determine reliable forecasts.

Commonly used and accessible weather forecast models include High-Resolution Rapid Refresh (HRRR), North American Model (NAM), National Blend of Models (NBM), Global Forecast System (GFS), and European Centre for Medium-Range Forecasts (ECMWF). These models not only forecast wind, but also temperature, pressure, humidity, precipitation, and other typical weather parameters. Each model varies in spatial resolution, forecast depth, and territory of forecast coverage, shown in Table 1.

Table 1: Parameters for each forecast model

Model	Spatial Resolution	Forecast Depth	Territory
HRRR	3km	1.5 days	North America
NAM	12km	2.5 days	North America
NBM	10km	11 days	North America
ECMWF	14km	10 days	Global
GFS	27km	10 days	Global

HRRR is designed for short-term forecasting with high-resolution and incorporates radar and satellite data to quickly update its forecasts. NAM is a mesoscale model that is used for regional forecasts and includes additional detail on terrain, allowing it to balance resolution and coverage over complex topographies. NBM blends multiple regional and global models together and is often considered more reliable by reducing individual model biases and uncertainty. GFS is a global model for medium to long-range forecasts so it can capture large scale weather patterns, but it is often not accurate at a local level. ECMWF is another global model and designed for medium-range forecasts and is highly regarded for its accuracy in data assimilation techniques [1].

1.2 Project Goal

The goal of this study is to perform a statistical comparison between the HRRR, NAM, NBM, GFS, and ECMWF wind forecasts. The accuracy of these forecasts will also be evaluated by comparing them with observed wind speed. The ultimate objective is to determine the predictive equation for wind speed using a regression analysis.

2 Methodology

2.1 Data Processing

To conduct this study, historical forecast and observed wind data needed to be collected. Python was utilized to manage the data collection and processing. All Python code used for data collection is shown in the Appendix. For downloading forecast data, the Herbie Python package was used, which is part of the NOAA Open Data Dissemination Program [2]. The downloaded data came in the GRIB format, which stands for "gridded binary" and is the international standard for spatial representation of meteorological data.

GRIB files for each of the five forecast models were downloaded at two forecast times—6-hour and 12-hour—from January 1, 2022, to December 31, 2023, at four specific times during the day: 00:00, 06:00, 12:00, and 18:00. This resulted in 14,600 files. All forecast models were fully represented, except for ECMWF, where data was only 96.7% complete. To ensure consistency, any dates and times with incomplete ECMWF data were removed from the dataset, resulting in a total of 12,521 GRIB files.

For the observed wind data, the Meteostat Python package was used, which sources data from NOAA [3]. The observed wind speeds from the T.F. Green weather station in Warwick, Rhode Island, were collected. Forecast data from each GRIB file was extracted at the exact coordinates of the weather station to ensure consistency in location when comparing forecasted and observed data. Figure 1 provides a visual representation of the GRIB data along with the T.F. Green weather station, illustrating the specific point at which wind speed values were extracted.

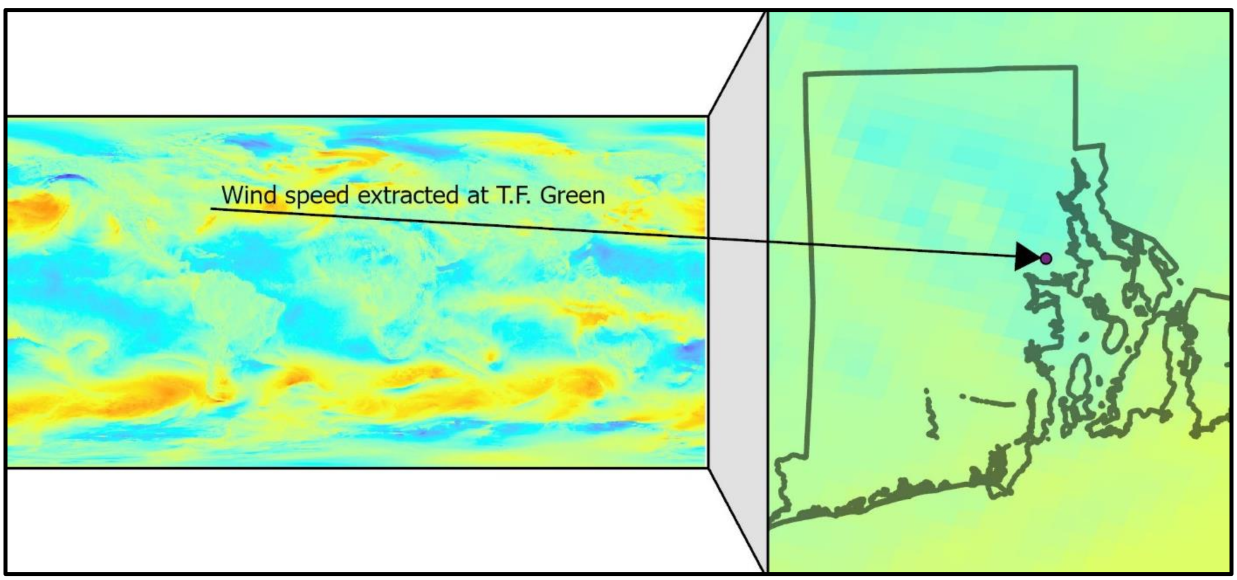

Figure 1: GRIB file and T.F. Green weather station point

2.2 Statistical Analyses

A series a statistical analyses will be conducted throughout this study. Each analysis has its own objective, attributing to the overall project goal. All tests will use a level of significance, $\alpha = 0.05$. The analyses and their objectives are as follows:

1. **Descriptive Statistics:** the objective of descriptive statistics is to represent the mean, median, standard deviation, and range of each of the five forecast models and the observed data. This

will enable a better understanding of the dataset's general characteristics. The distribution of the datasets will also be analyzed to determine whether the data is skewed, symmetrical, or follows a normal distribution.

2. **Paired T Test:** the objective of the paired T test is to compare the datasets to each other by determining if the mean of a dataset is equal to the mean of the compared dataset. Paired T tests are used when comparing paired data, and since the wind data represents the exact same location and time, it is considered to be paired. The tests will be conducted for each of the five forecast models to one another. Additionally, to compare the forecast model accuracy, a paired T test will be conducted comparing the observed data to each of the forecast models. These tests will be conducted for using both the population dataset (n = 2504) and a random sample dataset (n = 100) to determine how sample size impacts these tests. Table 2 depicts the null and alternative hypothesis as well as the test statistic for the paired T Test. The criteria for rejecting the null hypothesis is if $|T_0| > t_{\alpha/2, n-1}$, where $\alpha = 0.05$.

Table 2: Null Hypothesis and Test Statistic for the Paired T Test

$H_0: \mu_i - \mu_j = 0$ $H_1: \mu_i - \mu_j \neq 0$	for i, j = HRRR, NAM, NBM, GFS, ECMWF, and observed data
$T_0 = \dfrac{\bar{d}}{s_d/\sqrt{n}}$	where $\bar{d}$ = mean difference between the two datasets; s_d = standard deviation of the difference; n = sample size

3. **ANOVA Test:** the objective of the Analysis of Variance (ANOVA) test is to assess the differences in model response due to factors. In this study two model responses will be analyzed: wind speed and the forecast error. The forecast error is calculated by taking the absolute value of the difference between the observed wind speed and the forecasted wind speed. The factors that will be tested are both the forecast model and the forecast time, making this a two factor ANOVA test. There are five models, including HRRR, NAM, NBM, GFS, and ECMWF, and two forecast times, including 6-hour and 12-hour. The interaction between forecast model and forecast time will also be examined. A Tukey's test of comparison will also be conducted for each of the ANOVA tests to determine which significant factors are grouped together through a comparison of means. Tables 3 and 4 depict the hypotheses for both ANOVA tests. The criteria for rejecting the null hypothesis is if $F_0 > f_{\alpha, dfFactor, dfError}$, where $\alpha = 0.05$.

Table 3: Hypotheses for ANOVA test with Wind Speed as Response

H_0: The mean wind speed of each **Model**$_i$ are equal H_1: At least one **Model**$_i$ differs in mean wind speed	for i = HRRR, NAM, NBM, GFS, ECMWF
H_0: The mean wind speed of each **Forecast Time**$_j$ are equal H_1: At least one **Forecast Time**$_j$ differs in mean wind speed	for j = 6- and 12-hour
H_0: The mean wind speed of each **Model X Forecast Time**$_{ij}$ are equal H_1: At least one **Model X Forecast Time**$_{ij}$ differs in mean wind speed	

Table 4: Hypotheses for ANOVA test with Forecast Error as Response

H_0: The mean forecast error of each **Model**$_i$ are equal H_1: At least one **Model**$_i$ differs in mean forecast error	for i = HRRR, NAM, NBM, GFS, ECMWF
H_0: The mean forecast error of each **Forecast Time**$_j$ are equal H_1: At least one **Forecast Time**$_j$ differs in mean forecast error	for j = 6- and 12-hour
H_0: The mean forecast error of each **Model X Forecast Time**$_{ij}$ are equal H_1: At least one **Model X Forecast Time**$_{ij}$ differs in mean forecast error	

4. **Regression Analysis:** the objective of the regression analysis is to estimate the relationship between a dependent variable and one or more independent variables. In this study, the wind speed is the dependent variable, while each of the model forecasts are the independent variables. The coefficient of determination, R^2, provides a measure of how well the dependent variable is predicted by the independent variables. This analysis will undergo simple linear regressions to evaluate the predictive power of each of the forecast models at each forecast time, resulting in 10 regressions. Additionally, a multiple linear regression will also be conducted for each forecast time which predicts the dependent variable with each model forecast simultaneously. Tables 5 and 6 depict the regression equations and hypotheses for the simple linear regression and multiple linear regression.

Table 5: Regression equation and hypotheses for Simpe Linear Regression

$Wind\ Speed = \beta_0 + \beta_i X_i$	for i = HRRR, NAM, NBM, GFS, ECMWF
H_0: $\beta_i = 0$ H_1: $\beta_i \neq 0$	where β_i is the regression coefficient

Table 6: Regression equation and hypotheses for Multiple Linear Regression

$Wind\ Speed = \beta_0 + \beta_{HRRR}X_{HRRR} + \beta_{NAM}X_{NAM} + \beta_{NBM}X_{NBM} + \beta_{GFS}X_{GFS} + \beta_{ECMWF}X_{ECMWF}$
H_0: $\beta_{HRRR} = \beta_{NAM} = \beta_{NBM} = \beta_{GFS} = \beta_{ECMWF} = 0$ H_1: At least one $\beta_{HRRR}, \beta_{NAM}, \beta_{NBM}, \beta_{GFS}, \beta_{ECMWF} \neq 0$

2.3 Limitations and Assumptions

This study has several limitations that could affect the accuracy of the results. First, it was conducted at a single geographic location over a span of just two years. Given that weather patterns can vary significantly and continuously, a broader scope and a more extended time frame would likely improve the results. Another limitation is the discrete times of wind data collected being only at four specific times during the day (00:00, 06:00, 12:00, and 18:00), which could miss significant variations that occur outside these times. Continuous hourly monitoring would offer a more accurate representation of wind patterns throughout the day.

The statistical analyses used in this study depend on certain assumptions. Regarding the paired T tests, it is assumed that the datasets are inherently paired since the data is representative of the same time and location. The ANOVA test assumes the data is normally distributed despite the skewness of the data. However, the Central Limit Theorem can help address this assumption, in that the large sample size helps create a more normal distribution even though the evidence of skewness persists. The multiple linear regression assumes independence among the variables, even though they are considered paired in the T tests. The results of the T test indicates that the models are statistically different from each other, which supports the assumption of independence for the regression analysis.

3 Results

3.1 Descriptive Statistics

Table 7 presents the descriptive statistics for the wind forecasts at 6-hour and 12-hour intervals for each of the five forecast models, as well as for the observed data. The statistics include the sample size, mean, median, standard deviation, minimum, and maximum, with all values in knots.

Although these statistics do not inherently suggest any specific conclusions, some interesting patterns emerge. Notably, both the 6-hour and 12-hour forecasts follow a similar trend in their mean wind speeds, with NBM consistently showing the highest mean, followed by NAM, ECMWF, GFS, and HRRR in descending order.

Another observation is that the 12-hour forecasts tend to predict lower wind speeds compared to the 6-hour forecasts. This might suggest that longer forecast times are more conservative in predicting wind speeds. Additionally, the observed wind speed tends to be greater than all forecast models, with the exception of the NBM model, indicating that most forecast models tend to underestimate actual wind speeds.

Table 7: Descriptive Statistics for 6-hour / 12-hour forecasts and Observed wind speeds (knots)

Model	Sample Size	Mean	Median	Standard Deviation	Minimum	Maximum
HRRR	1255 / 1249	6.54 / 5.60	5.74 / 4.92	3.49 / 3.04	0.35 / 0.16	21.05 / 24.00
NAM	1255 / 1249	7.21/ 6.64	6.63 / 5.99	3.88 / 3.66	0.10 / 0.26	23.03 / 28.69
NBM	1255 / 1249	7.99 / 7.26	7.78 / 6.22	4.10 / 3.51	0.78 / 1.56	29.55 / 33.43
GFS	1255 / 1249	6.45 / 5.83	6.06 / 5.35	3.23 / 2.99	0.28 / 0.30	18.70 / 23.30
ECMWF	1255 / 1249	6.79 / 6.24	6.46 / 5.90	2.93 / 2.76	0.42 / 0.16	19.44 / 22.37
Observed	1255	7.44	7.02	4.64	0.00	26.03

To evaluate the distribution of wind data, the datasets were plotted as histograms and tested for normality. Wind data typically exhibits patterns that differ from a normal distribution due to the inherent variability in weather conditions. Specifically, wind data is often skewed to the left, indicating a higher frequency of lower wind speeds with fewer instances of higher wind speeds. This skewness was observed across all wind forecasts and is illustrated for the HRRR forecast in Figure 2. The histogram for HRRR displays a pronounced leftward skew, with the data falling well outside the 95% confidence interval for a normal distribution. The test for normality resulted in a p-value of < 0.005, providing strong evidence that the data is not normally distributed.

These findings were consistent across all forecast models and in the observed data, suggesting that standard assumptions of normality may not be applicable when dealing with wind forecasts.

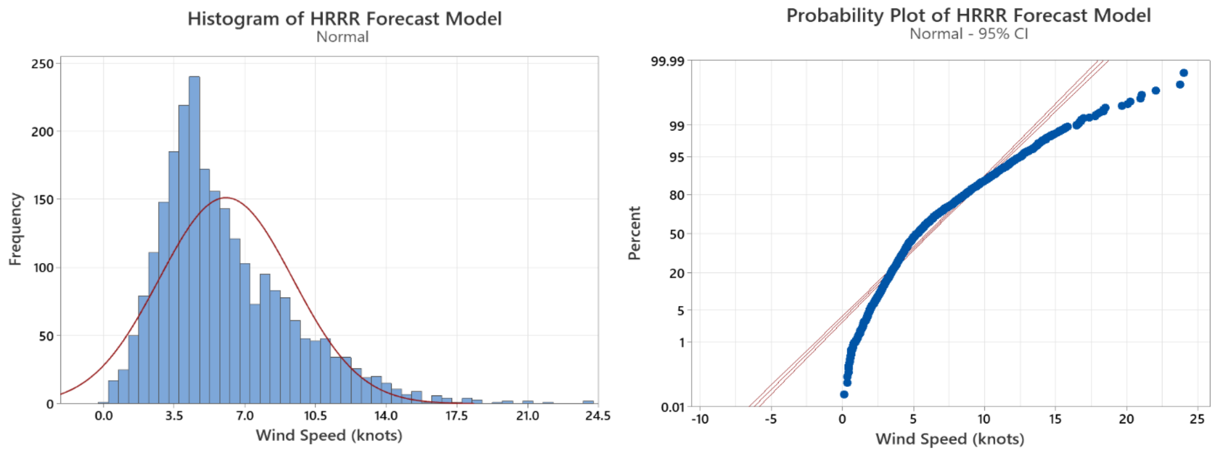

Figure 2: Histogram and Normal Probability plot for HRRR wind forecast data

3.2 Paired T Test

The paired T test is used to compare two related datasets to determine whether their means are statistically equivalent. In this analysis, a paired T test was applied each pair of forecast models to assess their differences. The initial tests on the full population datasets revealed that nearly all models were statistically different from one another, with very low p-values, most close to zero. The exception was the comparison between NAM and GFS, which had a p-value greater than the significance level ($\alpha = 0.05$), at 0.066.

This result aligns with the expectation that larger sample sizes increase the statistical power of a test but can also make even minor differences statistically significant. To further explore this, a random sample of 100 observations from the population datasets was taken conducted in paired T tests again. Although the t-scores were substantially lower with this smaller sample, most models remained significantly different, with low p-values. Notably, the NAM and GFS models still showed no significant difference, with a p-value of 0.582, indicating statistical equivalence. Additionally, the comparison between NAM and another model resulted in a marginal p-value of 0.056, suggesting a borderline significant difference. The results are summarized in Table 8, showing the t-scores and p-values from both the full population and the smaller sample T tests.

Table 8: Population and Sample Paired T Test results: T-scores and p-values

Population

	HRRR	NAM	NBM	GFS	ECMWF
HRRR	X				
NAM	21.20, 0.000	X			
NBM	41.44, 0.000	19.35, 0.000	X		
GFS	**1.84, 0.066**	19.82, 0.000	42.99, 0.000	X	
ECMWF	10.73, 0.000	9.07, 0.000	26.58, 0.000	12.23, 0.000	X

Sample

	HRRR	NAM	NBM	GFS	ECMWF
HRRR	X				
NAM	4.91, 0.000	X			
NBM	10.74, 0.000	3.26, 0.002	X		
GFS	**0.55, 0.583**	4.28, 0.000	9.91, 0.000	X	
ECMWF	2.96, 0.004	**1.93, 0.056**	6.07, 0.000	3.06, 0.003	X

A paired T test was also conducted comparing each forecast model to the observed wind data. The population results indicated that every model was statistically different from the observed data, with all p-values equal to zero. To explore the effect of a smaller sample size, a random sample of 100 was taken from the population datasets and the paired T tests were run again. The results of the sample T tests indicate that both NAM and NBM are statistically equivalent to the observed data, with p-values equal to 0.202 and 0.272, respectively. This suggests that these two models may be more consistent with observed wind patterns than other models. The results summarizing the T-scores and p-values for both the population and sample paired T tests are shown in the Table 9.

Table 9: Population and Sample T Test comparing forecast models to observed data

Model	Population T-score, p-value	Sample T-score, p-value
HRRR	8.72, 0.000	6.23, 0.000
NAM	9.98, 0.000	**1.28, 0.202**
NBM	8.47, 0.000	**1.10, 0.272**
GFS	9.26, 0.000	4.97, 0.000
ECMWF	9.38, 0.000	2.93, 0.004

3.3 ANOVA Test

An ANOVA test was performed to assess the impact of each forecast model, forecast time, and their interaction on the variability in model response. Two ANOVA tests were conducted: one where wind speed was the response variable, and another where the forecast error was the response variable.

The results of the first ANOVA, which examined wind speed, indicated that both model and forecast time were significant factors, with p-values equal to zero, suggesting that both factors strongly influence wind speed predictions. The interaction between model and forecast time, however, had a p-value of 0.228, indicating that the combined effect of these factors on wind speed was not statistically significant. Based on these results, we can reject the null hypothesis for both the model and forecast time, affirming that at least one model and forecast time differ from the others. However, we fail to reject the null hypothesis for the interaction, implying that this interaction does not significantly impact the variation in wind speed.

These findings align with the paired T test results, reinforcing the observation that the models vary from each other in terms of predicting wind speed. The ANOVA table summarizing the results is presented in Table 10.

Table 10: ANOVA Table for two-factor ANOVA test assessing differences in wind speed

Source	DF	SS	MS	F-Value	P-Value
Model	4	4126	1031.48	90.00	**0.000**
Forecast Time	1	1447	1447.40	126.29	**0.000**
Model X Forecast Time	4	65	16.15	1.41	0.228
Error	12510	143379	11.46		
Total	12519	149017			

A Tukey's test was conducted to evaluate how the forecast models and forecast times group together. The results showed that the GFS and HRRR models form one group, while all other models fall into separate groups. For forecast times, the 6-hour and 12-hour forecasts each form unique groups.

The second ANOVA test was conducted with forecast error as the response variable. The results indicate that model, forecast time, and their interaction are all significant factors, with p-values of 0, 0, and 0.016, respectively. This means we can reject all the null hypotheses for these factors, suggesting that each plays a significant role in forecast error. The ANOVA table summarizing the results is presented in Table 11.

Table 11: ANOVA Table for two-factor ANOVA test assessing differences in model error

Source	DF	SS	MS	F-Value	P-Value
Model	4	168.7	42.18	12.80	**0.000**
Forecast Time	1	60.2	60.17	18.25	**0.000**
Model X Forecast Time	4	40.4	10.11	3.07	**0.016**
Error	12510	41243.6	3.30		
Total	12519	41513.1			

A Tukey's test of comparison was conducted to identify groups of forecast models and forecast times that share similar characteristics. The results indicate that the NBM, NAM, and ECMWF models form one group, while the GFS and HRRR models form another distinct group. In terms of forecast times, the 6-hour and 12-hour forecasts each represent unique groups.

These findings suggest that while certain models share similar error patterns, others have distinctive behaviors. Additionally, forecast times can affect error patterns in ways that warrant separate grouping, indicating a potential impact on forecast accuracy based on the time frames.

3.4 Regression Analysis

Regression analysis was employed to understand the relationship between wind speed and forecasted wind speed, examining how accurately different models predict wind speed at various forecast times. A simple linear regression was performed for both 6-hour and 12-hour forecasts, and the results are presented in Table 12, showing the coefficient of determination (R^2) for each model at each forecast time.

The results indicate that the NAM model has the strongest predictive power when using a 12-hour forecast, achieving an R^2 of 61.28%. On the other hand, the NBM model delivers the highest accuracy with a 6-hour forecast, with an R^2 of 65.06%. Conversely, the ECMWF model demonstrates the lowest predictive accuracy, with an R^2 of 49.27% for the 12-hour forecast and 53.87% for the 6-hour forecast.

Table 12: Simple linear regression results

Model	**12-hour R^2**	**6-hour R^2**
HRRR	55.45%	63.92%
NAM	**61.28%**	61.27%
NBM	58.60%	**65.06%**
GFS	56.22%	60.81%
ECMWF	49.27%	53.87%

Figure 3 plots the relationship between NBM and observed wind speed with the fitted regression equation: **Y = 1.308 + 0.824 NBM.**

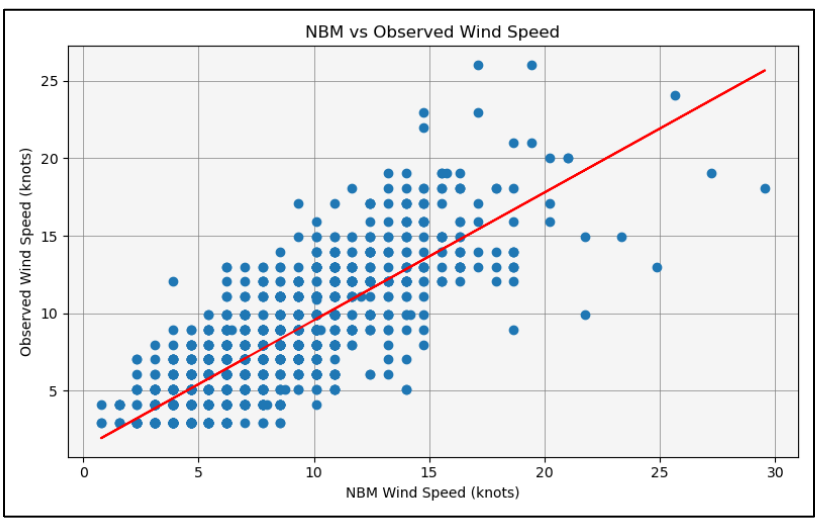

Figure 3: NBM versus observed wind speed with fitted regression equation

Given that the 6-hour forecast demonstrated higher accuracy compared to the 12-hour forecast, a multiple linear regression was conducted only for the 6-hour forecast. Initially, this analysis included all five forecast models as independent variables, however, the results showed that ECMWF was not a significant predictor, with a p-value of 0.129. Consequently, ECMWF was excluded from the regression model.

After removing ECMWF, the adjusted coefficient of determination (R^2-adjusted) for the revised multiple linear regression model was 69.88%, indicating a substantial improvement in predictive power compared to the simple linear regression models. The regression equation for the multiple linear regression, based on the remaining significant variables, is shown below:

$$Y = 0.830 + 0.327\ HRRR + 0.216\ NAM + 0.254\ NBM + 0.205\ GFS$$

The ANOVA table for the final multiple linear regression is shown in Table 13.

Table 13: Multiple linear regression results for 6-hour model forecasts

Source	DF	SS	MS	F-Value	P-Value
Regression	5	13297.0	332.2	653.5	**0.000**
HRRR	1	254.6	254.6	50.1	**0.000**
NAM	1	154.3	154.3	30.3	**0.000**
NBM	1	153.7	153.7	30.2	**0.000**
GFS	1	85.9	85.9	16.9	**0.000**
Error	1121				
Total	1125				

4 Conclusion

4.1 Key Findings

This study yielded several interesting insights into wind forecast models. The primary objectives were to statistically compare the five forecast models to each other and to observed wind data. The analysis revealed that, with the exception of GFS and HRRR, the five forecast models were significantly different from each other. However, the large sample sizes in the datasets made it challenging to demonstrate statistical equivalence among the forecasts. Despite these challenges, the variations between the models suggest that anyone analyzing wind forecasts should consider the model type rather than assume all models yield similar results. Additionally, the forecast times emerged as a significant factor, with the 6-hour forecasts proving more accurate than the 12-hour forecasts. Analyzing the model error, defined as the difference between forecasted and observed wind speeds, revealed that HRRR and GFS exhibited lower error compared to NAM, NBM, and ECMWF.

When evaluating the predictive power of each individual forecast model relative to observed wind speed, the NBM model at a 6-hour forecast time showed the highest accuracy, with an R^2 of 65.06%. This was followed by HRRR (63.92%), NAM (61.27%), GFS (60.81%), and ECMWF (53.87%).

Applying multiple linear regression improved these R^2 values to 69.88%, indicating an enhanced predictive power. Although this R^2 value is considered robust, there's potential for further improvement, suggesting a need for future work to refine these results.

4.2 Future Work

To build upon this study, additional data collection is required. First, historical forecast data for all models should be obtained from their origin, acknowledging that the datasets will have varying sample sizes. Additionally, a broader range of forecast times, from 1-hour to 10-day forecasts, should be included to evaluate the models' accuracy over different time frames.

Furthermore, wind forecast data from various locations would be beneficial to understand how geographical conditions affect forecast accuracy. Data collection should encompass diverse environments such as oceans, mountains, rivers, and various regions across the United States, provided observed wind data is available.

Exploring other meteorological parameters could also yield insights into factors impacting wind forecast accuracy. Variables such as wind direction, temperature, humidity, atmospheric pressure, and cloud cover might exhibit relationships with wind speed, influencing the reliability of forecasts. Including these parameters could guide more comprehensive future studies aimed at improving wind forecasting accuracy.

References

[1] National Weather Service. National Oceanic and Atmospheric Administration. Model Analyses and Guidance. https://mag.ncep.noaa.gov/model-guidance-model-area.php

[2] Blaylock, B. K. (2022). Herbie: Retrieve Numerical Weather Prediction Model Data (Version 2022.9.0) [Computer software]. https://doi.org/10.5281/zenodo.4567540

[3] Lamprecht, C. S. (2022) Meteostat (Version 1.2.0) [Computer software]. https://orcid.org/0000-0003-3301-2852

Appendix

```
import pygrib
import numpy as np
import os
import pandas as pd
import time
from herbie import FastHerbie
import multiprocessing
from meteostat import Point, Hourly, Daily
from datetime import datetime

# Downloading GRIB files using Herbie

# FastHerbie setup ##

models = ['nbm', 'hrrr', 'nam', 'ecmwf', 'gfs']
products = ['co', 'sfc', 'conusnest.hiresf', 'oper', 'pgrb2.0p25']
search_strings = ["10 m", "(?:U|V)GRD:10 m", "(?:U|V)GRD:10 m", ":10[u|v]:",
":[U|V]GRD:"]
times = ['00:00', '06:00', '12:00', '18:00']
hourly_freq = 24
periods = 365
fxx = [6, 12]
freq = str(hourly_freq)+'H'

for model, product, search_string in zip(models, products, search_strings):
    print('Starting model: ', model)
    start_time = time.time()
    for time1 in times:
        print('For time: ', time1)
        DATES = pd.date_range(start="2023-01-01 " + time1, periods=periods,
freq=freq)
        H = FastHerbie(DATES, model=model, product=product, fxx=fxx)
        H.download(search_string)
        end_time = time.time()
        print('Time taken: ', end_time - start_time)


# SORT GRIB FILES #

def ms_to_knots(ms):
    """
    Converts meters per second to knots
    """
    return ms * 1.94384
def wind_uv_to_dir(U, V):
    """
    Calculates the wind direction from the u and v component of wind.
    Takes into account the wind direction coordinates is different than the
    trig unit circle coordinate. If the wind direction is 360 then returns
zero
    (by %360)
```

```
    Inputs:
      U = west/east direction (wind from the west is positive, from the east
is negative)
      V = south/north direction (wind from the south is positive, from the
north is negative)
    """
    WDIR = (270 - np.rad2deg(np.arctan2(V, U))) % 360
    return WDIR

def wind_uv_to_spd(U, V):
    """
    Calculates the wind speed from the u and v wind components
    Inputs:
      U = west/east direction (wind from the west is positive, from the east
is negative)
      V = south/north direction (wind from the south is positive, from the
north is negative)

    Also converts speed from m/s to knots
    """
    WSPD = np.sqrt(np.square(U) + np.square(V))
    return ms_to_knots(WSPD)

def get_grb_indx(grb, lat, lon):
    """
    Function to get the index of the nearest grid point to a target latitude
and longitude
    """
    # Get the latitude and longitude grids
    lats, lons = grb.latlons()
    # Calculate the squared differences between the target latitude and each
latitude grid point
    lat_diff_sq = (lats - lat)**2
    # Calculate the squared differences between the target longitude and each
longitude grid point, considering globe
    lon_diff_sq = np.minimum((lons - lon)**2, (360 - (lons - lon))**2)
    # Calculate the squared distance from the target point to each grid point
    dist_sq = lat_diff_sq + lon_diff_sq
    # Find the index of the minimum squared distance
    min_idx = np.unravel_index(np.argmin(dist_sq), dist_sq.shape)
    return min_idx

def get_value_at_coord(grib_file, parameter_name_str, lat, lon):
    """
    Function to extract the value of a parameter at the nearest grid point to
a target latitude and longitude
    """
    try:
        # Open the GRIB2 file
        grbs = pygrib.open(grib_file)
        grb = grbs.select(name=parameter_name_str)[0]
        # Get the latitude and longitude grids
        lats, lons = grb.latlons()
```

```python
        # Run function to get grb index
        min_idx = get_grb_indx(grb, lat, lon)
        # Extract the data at the nearest grid point
        value = grb.values[min_idx]
        # Get the latitude and longitude at the nearest grid point
        nearest_lat = lats[min_idx]
        nearest_lon = lons[min_idx]
        return value
    except IndexError:
        # If no matches are found, return None
        return None

def get_date_time(file):
    """
    Function to get the date and time of the forecast of GRIB file
    """
    grbs = pygrib.open(file)
    grb = grbs.select()[0]
    date_time = grb.validDate
    return date_time

def get_forecast_time(file):
    """
    Function to get the forecast time of GRIB file
    """
    grbs = pygrib.open(file)
    grb = grbs.select()[0]
    forecast_time = grb.forecastTime
    return forecast_time


models = ['gfs', 'nam', 'ecmwf', 'hrrr', 'nbm']
def process_file(filepath, model, lat, lon):
    try:
        if model == 'nbm':
            wind_speed = ms_to_knots(get_value_at_coord(filepath, "10 metre
wind speed", lat, lon))
            wind_direction = get_value_at_coord(filepath, "10 metre wind
direction", lat, lon)
            date_time = get_date_time(filepath)
            forecast_time = get_forecast_time(filepath)
        else:
            u_value = get_value_at_coord(filepath, "10 metre U wind
component", lat, lon)
            v_value = get_value_at_coord(filepath, "10 metre V wind
component", lat, lon)
            if u_value is None or v_value is None:
                return None
            wind_speed = wind_uv_to_spd(u_value, v_value)
            wind_direction = wind_uv_to_dir(u_value, v_value)
            date_time = get_date_time(filepath)
            forecast_time = get_forecast_time(filepath)
        return date_time, forecast_time, wind_speed, wind_direction
```

```python
    except Exception as e:
        print(f"Error processing file {filepath}: {e}")
        return None

def sort_grib_files_parallel(model, lat, lon):
    path = os.path.join(r"C:\Users\flow\data", model)
    if not os.path.exists(path):
        print('Path does not exist:', model)
        return None

    files_list = []
    for root, dirs, files in os.walk(path):
        for file in files:
            filepath = os.path.join(root, file)
            files_list.append(filepath)

    pool = multiprocessing.Pool()
    results = []
    for filepath in files_list:
        results.append(pool.apply_async(process_file, (filepath, model, lat,
lon)))

    pool.close()
    pool.join()

    df = pd.DataFrame()
    dates = []
    forecast_times = []
    wind_speeds = []
    wind_directions = []

    for result in results:
        result_data = result.get()
        if result_data is None:
            continue
        date_time, forecast_time, wind_speed, wind_direction = result_data
        dates.append(date_time)
        forecast_times.append(forecast_time)
        wind_speeds.append(wind_speed)
        wind_directions.append(wind_direction)

    df['Date'] = dates
    df['Forecast Time'] = forecast_times
    df['Wind Speed'] = wind_speeds
    df['Wind Direction'] = wind_directions

    return df

if __name__ == '__main__':
    lat = 41.7235
    lon = -71.4270

    # Now you can call the sort_grib_files_parallel function for each model
```

```
    start_time = time.time()
    print('Sorting GFS Files...')
    df_gfs = sort_grib_files_parallel('gfs', lat, lon)
    df_gfs.to_excel('gfs_tfgreen_2023.xlsx', index=False)
    print('Time taken: ', time.time() - start_time)

    start_time = time.time()
    print('Sorting NAM Files...')
    df_nam = sort_grib_files_parallel('nam', lat, lon)
    df_nam.to_excel('nam_tfgreen_2023.xlsx', index=False)
    print('Time taken: ', time.time() - start_time)

    start_time = time.time()
    print('Sorting ECMWF Files...')
    df_ecmwf = sort_grib_files_parallel('ecmwf', lat, lon)
    df_ecmwf.to_excel('ecmwf_tfgreen_2023.xlsx', index=False)
    print('Time taken: ', time.time() - start_time)


    start_time = time.time()
    print('Sorting NBM Files...')
    df_nbm = sort_grib_files_parallel('nbm', lat, lon)
    df_nbm.to_excel('nbm_tfgreen_2023.xlsx', index=False)
    print('Time taken: ', time.time() - start_time)

    start_time = time.time()
    print('Sorting HRRR Files...')
    df_hrrr = sort_grib_files_parallel('hrrr', lat, lon)
    df_hrrr.to_excel('hrrr_tfgreen_2023.xlsx', index=False)
    print('Time taken: ', time.time() - start_time)


# Download Meteostat data #

# download data from meteostat
start = datetime(2022, 1, 1, 0, 0)
end = datetime(2023, 12, 31, 23, 59)

tfgreen_lat = 41.7235
tfgreen_lon = -71.4270

point = Point(tfgreen_lat, tfgreen_lon)
data = Hourly(point, start, end)
data = data.fetch()

winddata = data[['wspd', 'wdir']]
winddata = winddata.dropna()

# convert wind speed from km/hr to knots
winddata['wspd'] = winddata['wspd'].apply(lambda x: x /1.852)

winddata.to_excel('tfgreen_meteostat.xlsx', index=True)
```