Conference/Symposium: AMS 2024
Session: Operational and Local Meteorology

Evaluating Numerical Weather Prediction Forecasting Accuracy in Columbus, Ohio

Authors: Megan Shaffer, Sam Porter, Daniel Baltes, Annie Giovannucci
Acknowledgement: Dr. James DeGrand and Jackie Beck
Affiliation: The Ohio State University

Abstract

The use of numerical weather prediction models to generate daily weather forecasts is ubiquitous. Numerical weather prediction utilizes complex sets of equations and computer algorithms in order to simulate the state of the atmosphere at some point in the future. Accurate weather forecasts are essential to everyday life; operational meteorologists rely on them for daily forecasts, the power industry uses them to plan electricity generation, and many other uses. However, the complexity of the atmosphere can make it difficult to aggregate all inputs and produce a consistently accurate forecast. In order to ensure forecasts are representative of the atmosphere, it is important to look back and compare forecasts against measured data. Because calculations are only executed at specific points, it is not computationally feasible to have infinitely many points, creating challenges with model resolution. To acquire forecasts between points, the models utilize interpolation which can create room for error. This investigation will perform a case study within Columbus, Ohio by deploying a weather station equipped with a thermometer, tipping bucket rain gauge, and cup anemometer. This study will gather temperature, precipitation, and wind data at 00Z, 06Z, 12Z, and 18Z each day and compare the

data at the specified times to 3 major numerical weather prediction models: Global Forecasting System (GFS), North American Mesoscale (NAM), and High Resolution Rapid Refresh (HRRR). Through this process, the study will determine which forecast model is the most accurate to the true, measured observations made by our weather station. This will better enhance the forecasting skills of meteorologists and provide clarity on which model is most accurate in the Columbus region.

24-Hour Forecast Accuracy of Three Major Weather Prediction Models

1. Introduction

Forecasting takes on a variety of resources and weather models to predict a somewhat accurate weather forecast. However, oftentimes, the weather prediction models forecast an inaccurate value. Throughout this project, temperature, humidity, wind direction, and wind speed will be measured through our weather station for the following four times daily: 06Z, 12Z, 18Z, and 00Z. At concurring times, forecast data from three major weather prediction models from the previous day will be compared to the measured values of a variety of variables our weather station records. Through this process over the duration of a large dataset, calculating the deviation from the measured value for each time daily, meteorologists will come to a better consensus for which forecasting weather prediction model is most accurate in terms of the four forecasted variables.

Objectives

The goal of this study is to determine the most accurate NWP model for the Columbus, Ohio area. The objectives of this study are to (1) deploy a weather station, (2) measure temperature, humidity, wind, and precipitation, (3) collect corresponding forecast data from 3 NWP models, and (4) compare both datasets to find trends and come to a data backed conclusion.

Background

There are many forecasting models that exist today to help meteorologists make predictions. The three major models used in the United States are the Global Forecasting System (GFS), North American Mesoscale (NAM), and High Resolution Rapid Refresh (HRRR). The complexity of the atmosphere can make it difficult to aggregate all inputs and produce a

consistently accurate forecast. Excluding severe weather, the most important measurements of the atmosphere include temperature, humidity, wind direction, and wind speed. Recording the forecast of the three major weather models 24 hours before taking measurements in the field and comparing the results will yield the accuracy of the models. Testing accuracy of weather prediction models has been done before, such as Uden et al. (2023) putting four winter precipitation models to the test in the Rocky Mountains. This was in effort to aid airports in knowing which models can best predict when to shut down during a snowstorm. Uden et al. (2023) used probability thresholds, while this study will use exact measurements to determine accuracy. Adams-Selin et al. (2023) put one model to the test over multiple years. They tested hail forecast accuracy through the eyes of a volunteer sample of the public. Participants were asked to evaluate the forecast of hail after every storm from 2019-2021. Adams-Selin et al. (2023) found that the accuracy increased over the three year study. Similarly, Giordano et al. (2013) conducted a study comparing the WRF model to in-situ measurements based out of The Observatorio del Roque de los Muchachos and concluded that the WRF was in good agreement for that region. Building off of these prior investigations, this study tests temperature, humidity, wind speed and direction, and rainfall. Meteorologists around the world need to know the most accurate models for predicting atmospheric conditions in order to communicate the forecast to the general public as well as industries that keep the world running.

2. Data & Methods

This study used Loggernet software to program the data logger. In this software, a program collected temperature, humidity, wind and precipitation data. The temperature probe collected maximum, minimum, and average temperature for every half hour. Code similar to this

was used to collect maximum, minimum, and average wind speed for every half hour. Despite the code collecting every half hour, the model data were only collected every 6 hours at 00z, 06z, 12z, and 18z. This is because models do not produce a prediction as resolute as every half hour but instead produce 24 hour forecasts with 3 hour resolution. These and similar codes to collect precipitation data were sent to a data logger. A flux tower was assembled and deployed as the weather station for data collection at Waterman Farm. A temperature probe and anemometer were fixed to the tower cross bar with an additional small tripod equipped with a tipping bucket rain gauge. The tower also held an enclosure, encasing instrumentation wiring and connecting it to the data logger. Initially, data were collected from October 6th, 2023 to October 26th, 2023; however, there was an issue with the data logger and all data from that period was unrecoverable. Because of this, a smaller tripod equipped with all instrumentation previously described was deployed and collected data from November 4th, 2023 to November 21st, 2023. Figure 1 shows the tripod setup from the successful data collection campaign.

Figure 1: Tripod Equipped with Temperature/Humidity Sensor, Anemometer, Tipping Bucket Rain Gauge, and Data Enclosure

In addition to this station, seven other stations around Columbus operated by the Ohio State University collected the same data. In order to get a representative sample of Columbus, a study of land cover use was performed to then be averaged proportionally with our collected data. Figure 2 below shows the National Land Cover Database for the city of Columbus with stars representing Ohio State weather stations and a triangle marking our team's tripod.

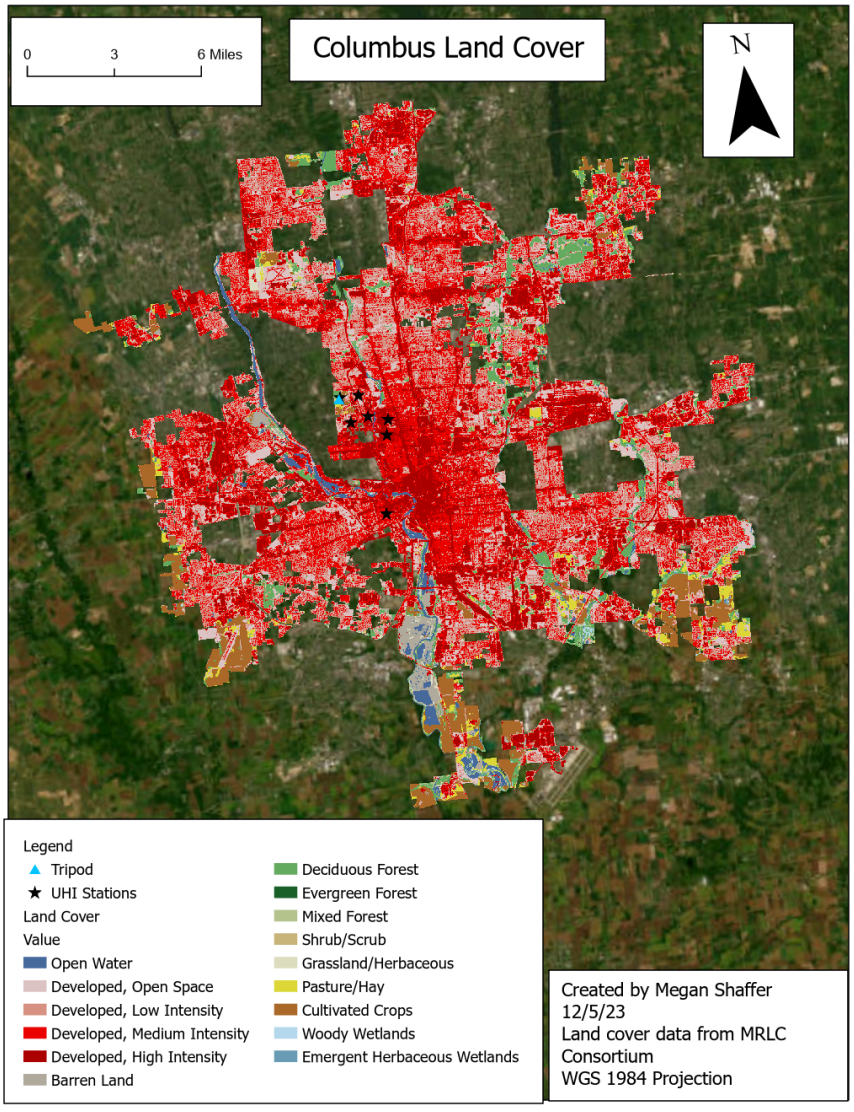

Figure 2: Columbus Land Cover Use with Weather Collection Locations

With the land cover map, an analysis of land cover was performed and is summarized in Figure 3. As shown, Columbus, consists primarily of developed land of different intensities. 21%

of Columbus land use classifications did not have stations. The amount of stations per represented land use category is shown on the right in Figure 3. Using this data, the stations were scaled to be proportional to Columbus's land cover. Because stations only covered 79% of Columbus's land cover types, that 79% total was scaled to 100% to be able to multiply the station values appropriately. Any data referred to as measured data in later analyses refers to this combination of campus weather stations with our weather station proportional to land use. Future studies could deploy more stations making sure to cover every land cover classification.

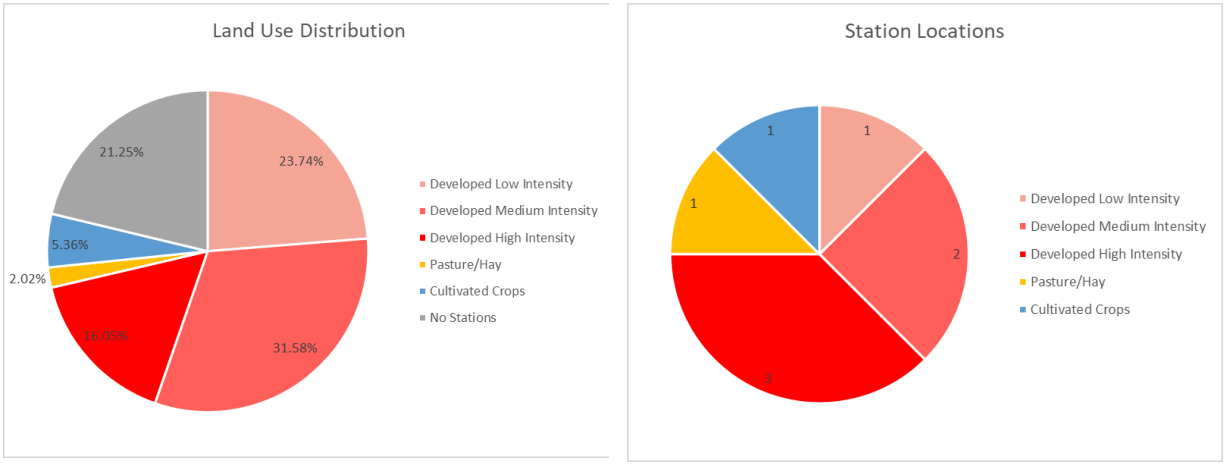

Figure 3: Land Use Distribution of Columbus (Left) and Station Location Across Different Land Use Classifications (Right)

Once the data were collected and formatted, we began analyses by creating time series of our variables of interest: temperature, dewpoint, and wind speed. Because the models report temperature and dewpoint temperature in degrees Fahrenheit (℉), all data were analyzed using units of Fahrenheit. Although Celsius is a universal temperature unit, Fahrenheit is accepted as the standard in the United States and the broad public has a reference for what ℉, so it was our decision to keep the data representations in a reference scale that is well understood in the study region. The wind speed was measured and reported in m/s. Additionally, the HC2S3 temperature

and humidity sensor has an uncertainty of ±0.1 ℃ which equates to ±0.18 ℉. The 014A anemometer used has an uncertainty of 0.11 m/s.

Another analysis performed was for error. In this case, the measured value for each variable was subtracted from each of the models in order to see a time series of error. Lastly, scatterplots for each of the 3 variables were created to further evaluate the direction in which the models varied from measured data. After that, standard error was calculated followed by r correlations to better quantify and conclude which model was most accurate for each variable.

Although precipitation amounts were recorded throughout the study period, there was very little precipitation. The models only had 6 instances of precipitation predictions which provides such a small dataset that any conclusions drawn from it would be flawed in their experimental design. Because of this, no analyses of precipitation were performed for this study. This could be an interesting variable to analyze in future studies if the area of interest has enough precipitation to create a large sample size.

3. Results and Discussion

Time Series Analyses

Figure 4 shows the time series of temperature through the study period. The model data are generally consistent with one another. The measured data are consistently either above or below the predicted temperature for the models. Actual high and low temperatures occurred later than what model predictions represent. This could be due to the limited collection of model data at 06Z, 12Z, 18Z, and 00Z. Comparing the amplitudes, it can be concluded that the models predicted a lower high and low temperature than what was measured. It can be concluded that the high temperature occurs later than 18Z (1pm EST) and the low later than 6Z (1am EST). The

expedited model data poses relative difficulty in determining which model is most accurate compared to what was measured. Comparing the amplitudes of the model data, the highs look consistent. The NAM and the HRRR seem to dramatically underestimate the low temperature. The GFS seems to dramatically overestimate the low temperature. The low temperature can be influenced by outside factors such as wind speed and cloud cover. This could explain the variation in the predicted model values. Data regarding cloud cover was not collected, so conclusions on model accuracy as a function of cloud cover cannot be drawn. Cloud cover could be an interesting component to integrate into further studies.

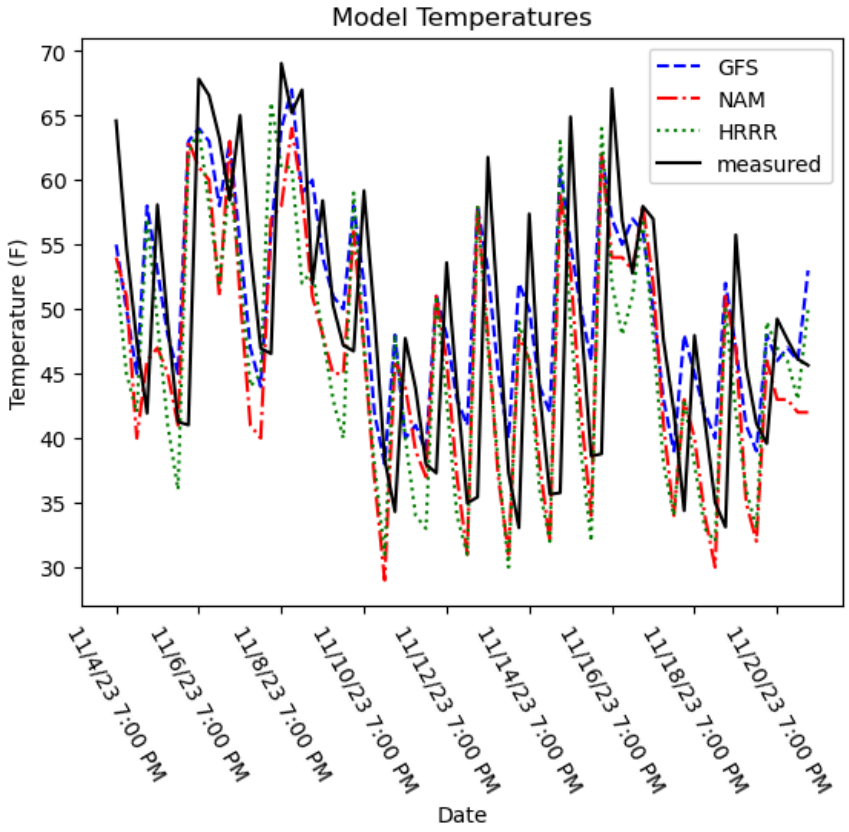

Figure 4: Model Predicted Temperatures Compared to Measured Temperatures

In Figure 5, the model forecast data is represented as a difference in temperature from the measurements taken in the field. With this representation, data greater than 0 means the model predicted greater than the measured, and data less than 0 means the model predicted less than measured. The HRRR model is seen to have the highest volatility, overestimating and underestimating the forecasted temperature more than the other models. The NAM model is the most consistent as the forecasted temperature is usually closest to the measured temperature. The magnitude of underestimation for the GFS model was smaller than for the other two models, but it still overestimated temperature with similar magnitude compared to the HRRR model. In all three models, the overestimation of temperature is of higher magnitude than the underestimation of temperature; overestimations reach almost 30℉ around 11/15 while underestimations do not even reach 20℉. The GFS model had an average error of 6.87 degrees F, the NAM model had an average error of 8.40 degrees F, and the HRRR model had an average error of 9.67 degrees F. In conclusion, the GFS model is the most accurate for forecasting 24 hour temperature. The GFS model typically does well forecasting climatically normal temperatures without a lot of convection happening. During the time period studied, the temperatures were normal for the study area and had very little convection, so the results align with the weather conditions.

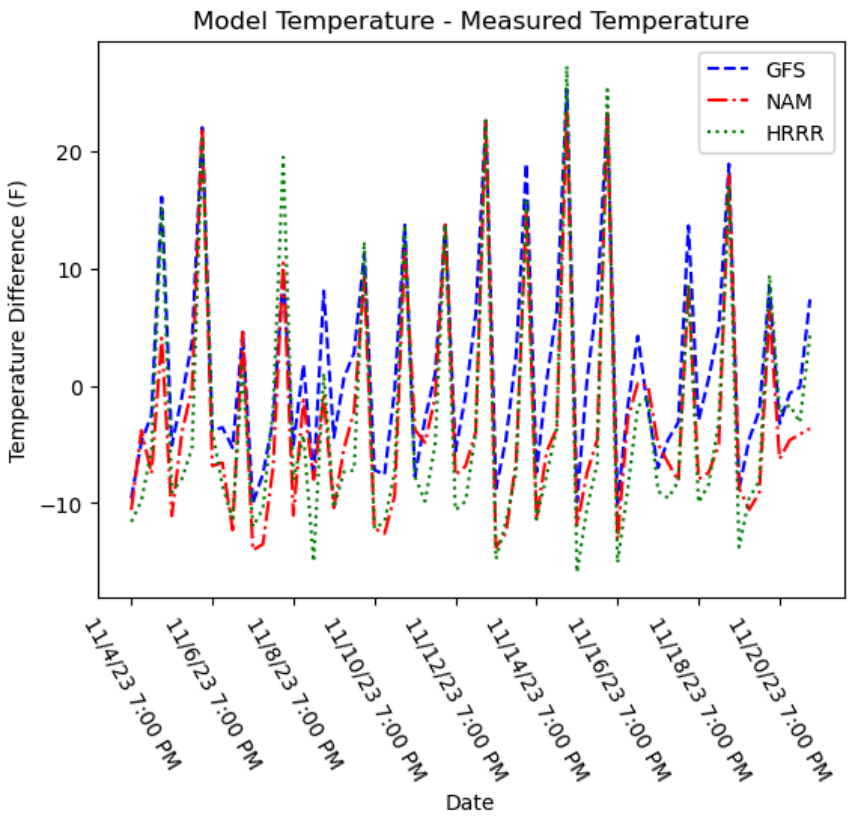

Figure 5: Model Temperature Error

In Figure 6, dewpoint forecast temperatures from the three models are compared to the measured values of dewpoint 24 hours after the model forecast data were collected. In Figure 7, the model dewpoint temperatures are represented as an error from the values measured in the field. Both figures show that the NAM model overestimated dewpoints more consistently than the other models, while the GFS model consistently underestimated dewpoint temperatures. The HRRR model was once again the most volatile, claiming the highest over and underestimations of dewpoint. Although the HRRR was the most extreme, it did report a lot of errors close to zero, and as a result, it was not the most inaccurate in forecasting dewpoint temperatures. The NAM model was the most accurate with an average error of 3.41℉, the HRRR model had an average error of 4.93℉, and the GFS model had an average error of 5.19℉. In conclusion, the NAM model was the most accurate in forecasting 24 hour dewpoint temperatures.

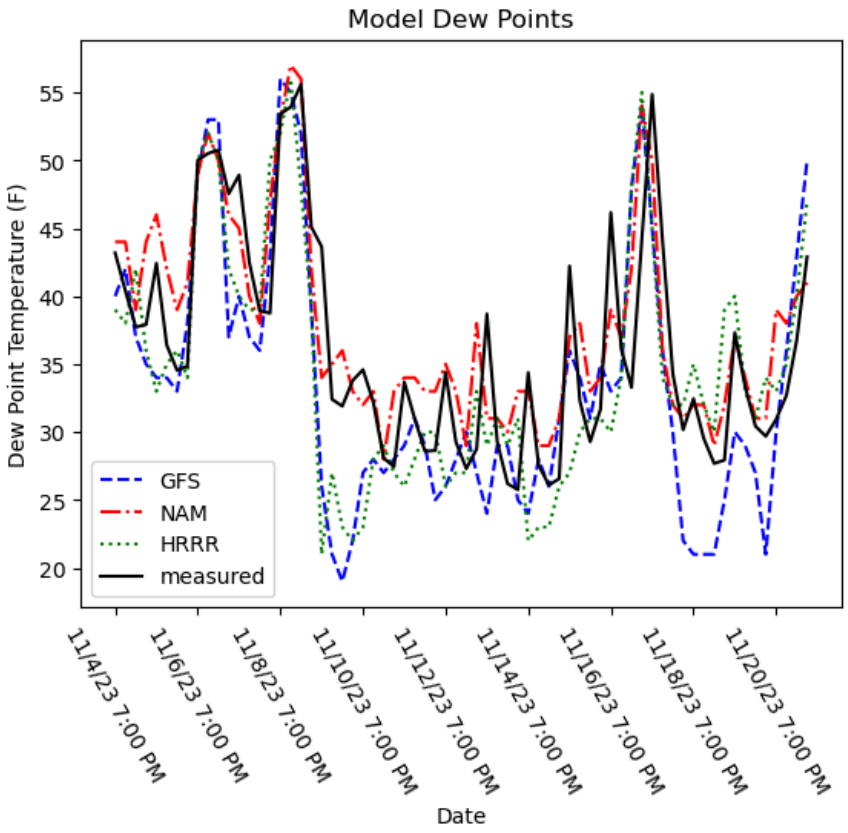

Figure 6: Model Predicted Temperatures Compared to Measured Temperatures

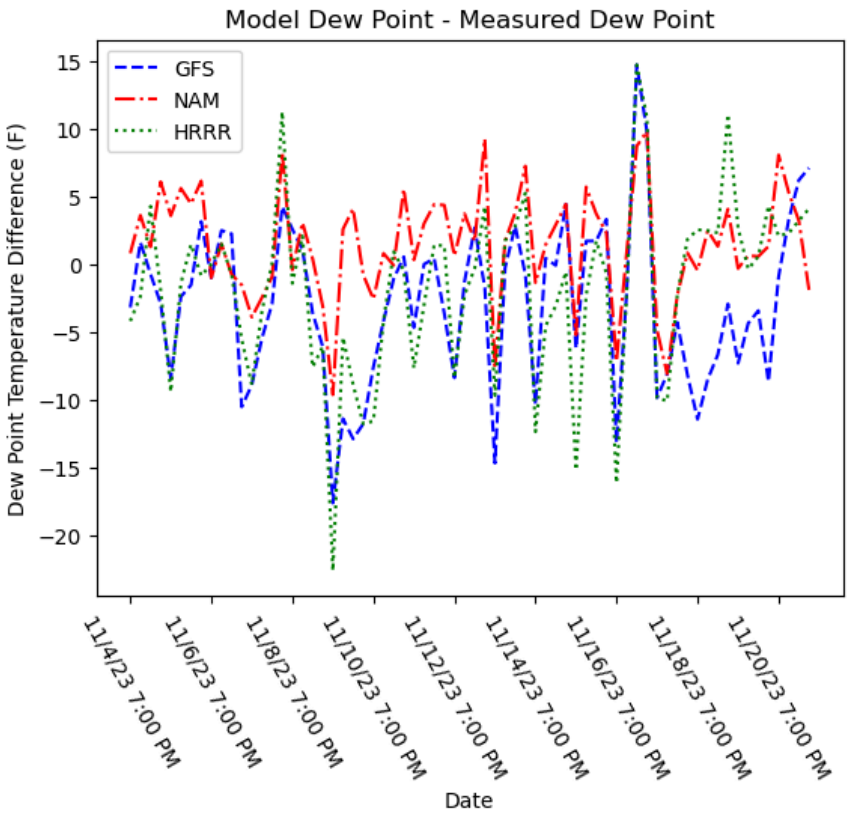

Figure 7: Model dewpoint Temperature Error

Figure 8 is a line graph of the comparison between all models of interest and the measured data for wind speed. This is a representation of sustained winds over the study duration. The GFS, NAM, and HRRR all consistently overestimate the wind speed. The NAM overestimated the wind speed by the greatest amount compared to the others.

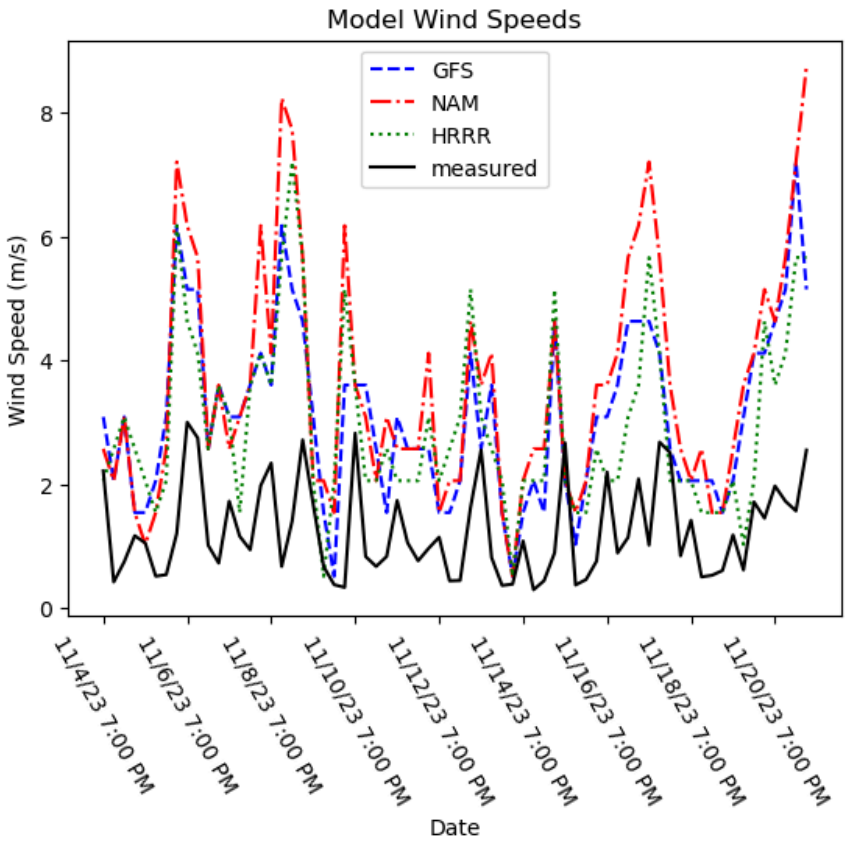

Figure 8: Model Predicted Wind Speeds Compared to Measured Wind Speeds

Figure 9 is a line graph of the difference between the models' predicted wind speed with the measured values. The NAM and the HRRR are most consistent with one another, while the GFS seems to predict higher values. It can be concluded that models were timely in predicting the max and low wind speeds, as there exists very little variation between the positioning of the amplitudes of each model.

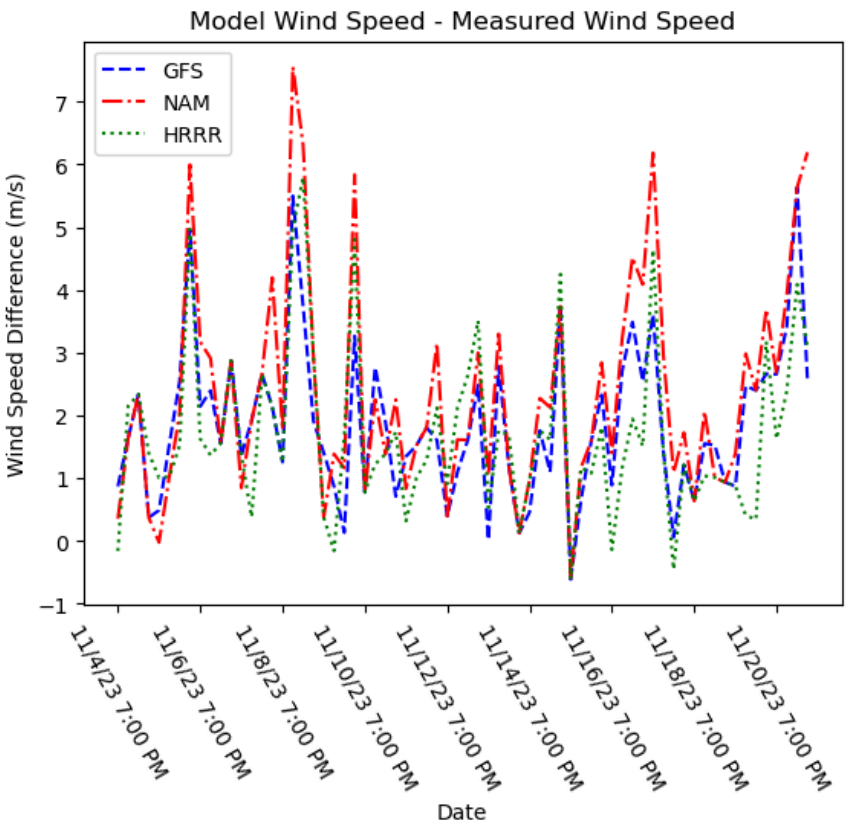

Figure 9: Model Wind Speed Error

Scatterplot Visualizations

To further visualize deviations between measured data and modeled data, scatterplots were created for each of the three variables. Each scatterplot shown in Figures 10-12 depicts the model variable on the y-axis and the measured variable on the x-axis. If the models perfectly matched the measured data, there would be a slope of 1 and an intercept of zero. Points above that ideal line mean that the model ran high whereas points below that ideal line mean that the model ran low. Figure 10 provides some interesting insight on the temperature results. It appears that all of the models drastically overestimated temperature at lower measured temperatures

more often than at higher measured temperatures. It appears for all the models, they primarily ran higher than the measured and often did not underpredict temperature.

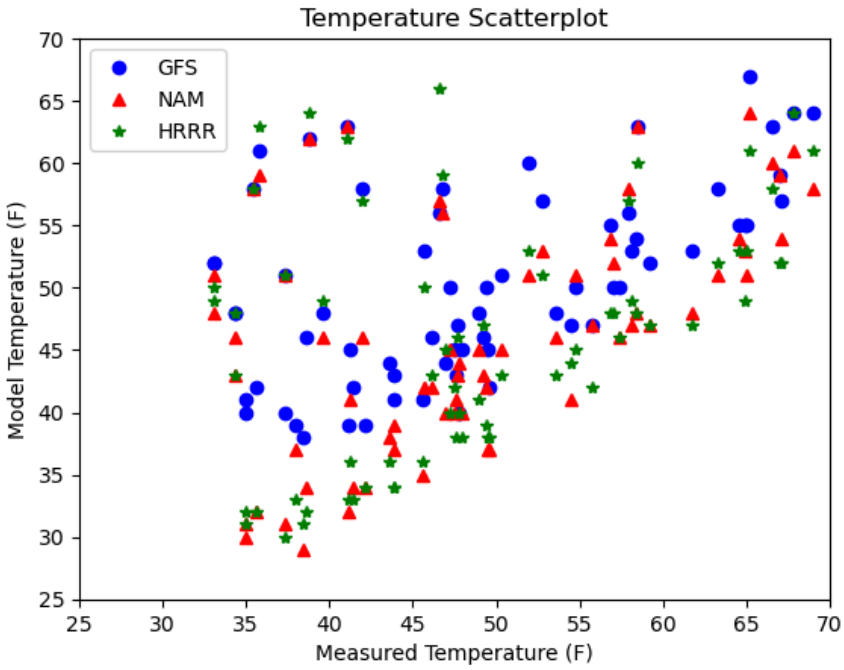

Figure 10: Temperature Scatterplot for Each Model

Shifting to dewpoints, Figure 11 shows that the GFS and HRRR tended to undermodel the dewpoint at lower temperatures whereas NAM instead slightly overmodeled at low and mid dewpoints.

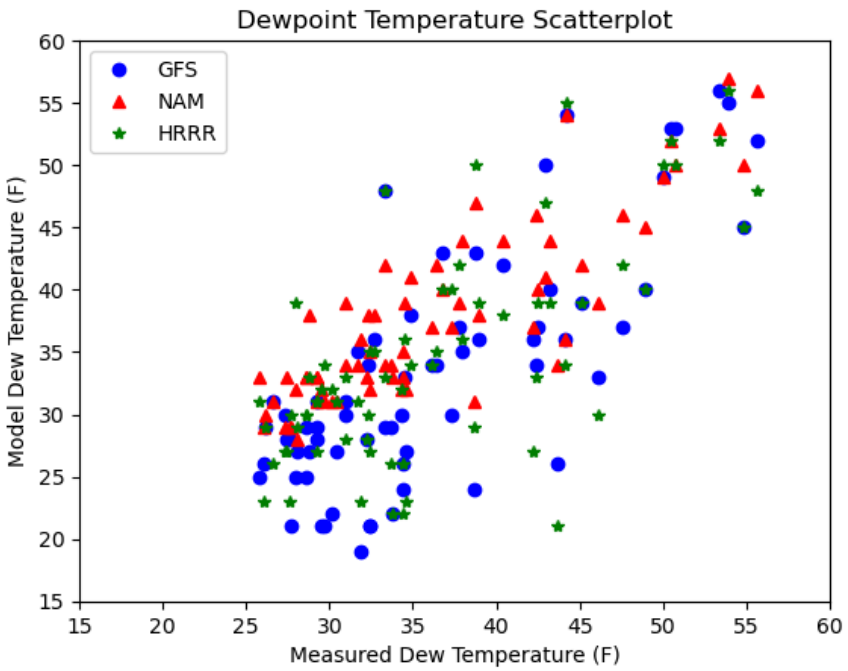

Figure 11: Dewpoint Temperature Scatterplot for Each Model

Looking last at wind speed, Figure 12 shows a large difference as depicted in prior time series between model data and measured data. The model data has a much larger range from close to 0 m/s all the way to 8 m/s, whereas the measured data only reaches a maximum of approximately 3 m/s. This large discrepancy in wind is an unfortunate downside to this project's data collection. The 3 models used in this study predict wind speeds at 10m above ground level whereas the weather stations used for Columbus measured data measured wind at around 2m above the ground. Winds increase in speed as height increases, so it is expected that these two measurement heights would have different wind speeds. The measurements taken can include more surface roughness factors and not be high enough to be unobstructed by structures and trees, but the modeled data at 10m evades many of those roughness elements. Future studies could be done to create a wind profile from measured data and extrapolate to 10m. Greater

investment could also be used to deploy meteorological stations with anemometers at 10m to acquire a more even comparison.

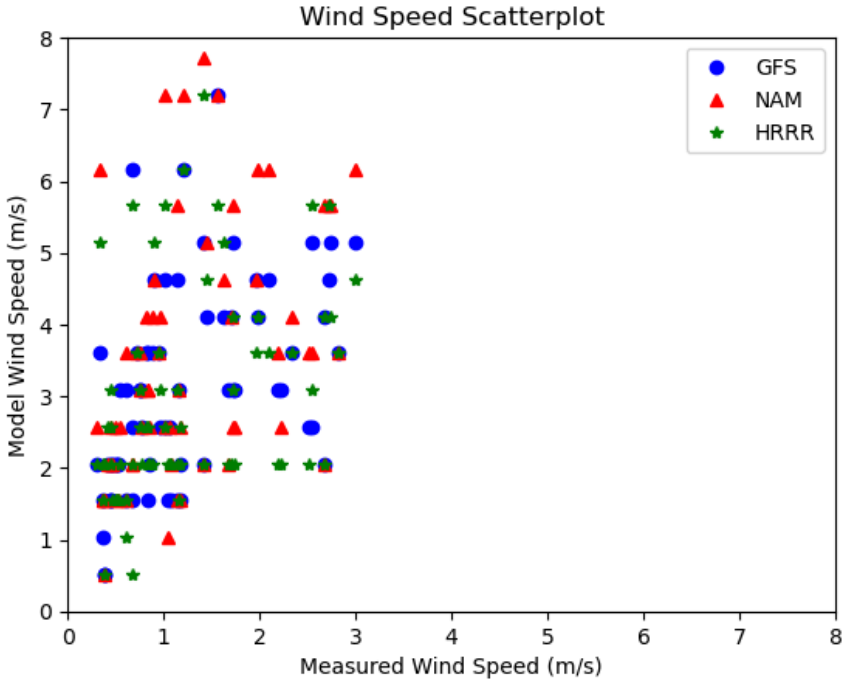

Figure 12: Wind Speed Scatterplot for Each Model

Analyses of Error

Table 2 shows the absolute error for the parameters of temperature, dewpoint, and wind speed between the measured values and model values. For temperature, the HRRR had the highest absolute error at 9.67, while the GFS had the lowest at 6.87. For dewpoint, the GFS had the highest absolute error at 5.19, while the NAM had the lowest at 3.41. For wind speed, the NAM had the highest absolute error at 2.38, while the HRRR had the lowest at 1.75. It can be concluded that there is immense variation between the models' absolute error for each studied parameter. In regards to Table 2, it can be concluded that for temperature, the GFS was more accurate due to the smallest average absolute error. In terms of the dewpoint, the NAM was more

accurate while the HRRR was more accurate for wind speed. The most accurate model for each variable is highlighted in green in table 2 below.

Table 2: Average Absolute Error for Each Variable of the Three Models

	GFS	NAM	HRRR
Temperature (℉)	*6.87±0.18*	*8.40±0.18*	*9.67±0.18*
Dewpoint (℉)	*5.19±0.18*	*3.41±0.18*	*4.93±0.18*
Wind Speed (m/s)	*1.88±0.11*	*2.38±0.11*	*1.75±0.11*

Table 3 shows the Pearson correlations (r) for each model and parameter. An r value of 1 means that there is a perfectly linear relationship between the values being compared. Each cell in this table compares whatever model and variable to the corresponding measured variable. As shown, the r values have a wide range. The strongest Pearson correlations for each variable are highlighted in green.

Table 3: Pearson correlations

	GFS	NAM	HRRR
Temperature	*.521*	*.550*	*.414*
Dewpoint	*.776*	*.594*	*.693*
Wind Speed	*.474*	*.555*	*.379*

4. Conclusion

This study, in determining the accuracy of three major NWP Models, displays varying results towards the accuracy of temperature, dewpoint, and wind speed. Through the analysis of NWP Models, this study presents evidence that the most accurate model when forecasting temperature is GFS. Meanwhile, when forecasting dewpoint temperatures, the NAM is the most

accurate. In terms of the wind speeds, the HRRR model is the most accurate to the measured data. Table 4 displays the rankings each NWP model upholds in terms of temperature, dewpoint, and wind speeds throughout the study period.

Table 4: Rankings of NWP Models

	Most Accurate	→	**Least Accurate**
Temperature (F)	*GFS*	*NAM*	*HRRR*
Dewpoint (F)	*NAM*	*HRRR*	*GFS*
Wind Speed (m/s)	*HRRR*	*GFS*	*NAM*

Through better understanding the accuracy of NWP Models, meteorologists in all sectors of the field can achieve a stronger confidence and clarity in understanding the accuracy of GFS, NAM, and HRRR. With forecasts delivered multiple times by meteorologists on a daily basis, this study improves the accuracy of the forecasts, which allows for the general public to possess more accurate weather forecasts.With more accurate forecasts, the public's confidence in forecasts increases, therefore, providing safety in times that severe weather strikes.

This study is relevant to the scientific community and similar to Jaseena et. al (2020), who studied the machine learning aspect of weather models to climatologically determine their accuracy. This study stated the importance of accurate forecasting to allow for successful industry operations such as: agriculture, tourism, airport systems, etc.

For future research on model forecasting accuracy, improvements can be made. To better estimate the accuracy of 24-hour model forecasting in Columbus, more stations would need to be deployed in more diverse land cover zones. This project had stations mainly north of the downtown area; ideally more consistent spacing around the metro area would yield more accurate conditions for the environment in Columbus. Additionally, most climate models

measure sustained wind speed at 10 meter elevation heights from the surface. Setting up stations that also measure wind speed at that height would correlate wind speeds more accurately. In addition, using more weather models in the research would strengthen the results in determining the most accurate model. This experiment was only conducted during two weeks out of the year, but conducting this experiment throughout different seasons could potentially yield different results. Due to the two week period of collecting data exhibiting drought conditions, collecting data over more than two weeks would also allow for more precipitation data to be collected and analyzed. With the evidence provided in determining the accuracy of NWP Models, this study allows us to draw generalized conclusions that GFS is most accurate for forecasting temperature, NAM is most accurate for forecasting dewpoint temperatures, and HRRR is most accurate for forecasting wind speeds during this time period in Columbus, Ohio.

References

Adams-Selin, R. D., and Coauthors, 2023: Just What Is "Good"? Musings on Hail Forecast Verification through Evaluation of FV3-HAILCAST Hail Forecasts. *Wea. Forecasting*, **38**, 371–387, https://doi.org/10.1175/WAF-D-22-0087.1.

Giordano, C., and Coauthors, 2013: Atmospheric and Seeing Forecast: WRF Model Validation with In-Situ Measurements at ORM. *Mon. Not. Roy. Astron. Soc.*, **430**, 3102-3111, doi:10.1093/mnras/stt117.

JaseenaK., U and Binsu C. Kovoor. "Deterministic weather forecasting models based on intelligent predictors: A survey" *J. King Saud Univ. Comput. Inf. Sci.* 34 (2020): 3393-3412.

Uden, D. M., M. S. Wandishin, P. Schlatter, and M. Kraus, 2023: Evaluation of Probabilistic Snow Forecasts for Winter Weather Operations at Intermountain West Airports. *Wea. Forecasting*, **38**, 1341–1362, https://doi.org/10.1175/WAF-D-22-0170.1.