The Utilization of Linear Regression Models for Weather Prediction

Michael Bell

Kennesaw Mountain High School

Advanced Scientific Research IV

Amanda Dennis

4/21/2024

Abstract

Machine learning with linear regression is a robust process that allows relationship-driven predictions for multiple variables. This study utilized the engineering design process and quantitative data collection to analyze the accuracy of linear regression for weather prediction and create a model that allows users to generate their own weather forecasts up to a year in advance with corresponding errors. The dataset for this study consisted of daily summaries from the Atlanta Hartsfield-Jackson Airport from the dates 1/1/1990 to 3/3/2024. Using this dataset combined with a Python package, Scikit-learn, I created a linear regression model to predict the weather over multiple 10-day periods to analyze its accuracy. The results of my 32-sample forecast study with the outlier removed yielded a mean error of 3.060511°F, a median of 2.875193 °F, a range of 5.046103 °F, a standard deviation of 1.286090 °F, a r-squared correlation of 0.902931, and a variance of 1.654026 °F. In conclusion, linear regression and the web application created in this study can benefit numerous societal fields, such as travel, business, public safety, agriculture, and energy production. While the web application allows users to create forecasts a year in advance, it is recommended that they stay as precise as possible to their desired date as possible for minimal error. I also recommend a future study evaluating the trend between possible error and date range to find if there is a method to increase the forecast range of the regression model without significantly compromising forecast confidence.

Acknowledgments

I would like to thank Amanda Dennis, Dr. Eden Hunt, my mentor Bry Burnham, and Cathy Stump, who all contributed to the development, editing, and fulfillment of this study. The time and effort these individuals spent assisting in this engineering process, as well as research development, heavily contributed to the collection of skills and knowledge needed to make this study a success. This level of research and technological innovation would not be possible without their steady guidance and constructive feedback — thank you.

Table of Contents

The Utilization of Linear Regression Models for Weather Prediction

Chapter 1: Introduction

Mark Twain once said, “Climate is what we expect [;] weather is what we get” (Twain, 1887). Weather forecasting, which is often inaccurate and unreliable, especially proves this phrase true. According to the National Oceanic and Atmospheric Administration (NOAA, 2024), a seven-day forecast is only correct about 80% of the time, and for a 10-day forecast, this accuracy falls to 50%. This uncertainty can produce many implications for society (Cutler et al., 2022; Jaseena & Kovoor, 2022; Kumar & Nabi, 2019). To solve this issue, researchers are looking to a promising method called linear regression, which uses artificial intelligence and machine learning algorithms to predict atmospheric phenomena based on vast amounts of training data (Kumar et al., 2017). The question, however, is, are they more accurate? Bhandari et al. (2023) conducted a previous study that proposed a weather prediction system based on linear regression, and they found that current prediction methods are usually more correct than those of their linear regression model.

While their result could be more impressive, their research needs various elements to validate such a conclusion. They do not specify the parameters used to train the model, the regression model used, the data the model would output, and the error between the model's predictions and actual weather. This level of uncertainty makes their conclusion hardly supported. An additional study done by Baral et al. (2020) has a similar goal, but their study is more specific in its methodology. They used the parameters of year, month, day, max temperature, and minimum temperature from 2009 to 2019 from the Kathmandu Airport station to develop their prediction model to forecast the next day's minimum and maximum temperature. Unfortunately, this study only tests for the temperature of the next day, and as stated previously, the issues in today's forecasts originate from predicting larger time intervals, such as 10-day forecasts (NOAA, 2024). My research addressed previous studies' limitations and evaluated linear regression accuracy over a 10-day period through the engineering design process. My research examines not only the efficacy of linear regression but also the issues with current forecasts, the evolution of forecasting, the importance of weather forecasting, and the process of forecasting. Understanding these elements of weather prediction is crucial to improving forecasting techniques and solving the problem of unreliable forecasting (Cutler et al., 2022; National Centre for Atmospheric Science [NCAS], 2023).

Statement of the Problem

Incorrect weather reports can affect travel plans and activities, impact the economic supply and demand by damaging infrastructure, and reduce trust in news organizations, leading to less preparation for severe weather hazards (Cutler et al., 2022; Jaseena & Kovoor, 2022; Kumar & Nabi, 2019). For example, 700,000 Europeans died from an unforeseen heatwave in 2007, where meteorologists inaccurately predicted the maximum temperatures (Fink & Lemburg, 2022). In addition to affecting safety, overprediction and underprediction of wind conditions lead to unnecessary costs and increased air pollution (Cutler et al., 2022). Unreliable weather forecast effects also extend to the agriculture sector as sudden, unpredicted freezes create extensive crop destruction and revenue loss (Cutler et al., 2022). The importance of weather forecasting scales beyond public safety, energy production, and agriculture. It also impacts trade, aviation, and travel; therefore, accurate weather prediction is a vital resource for many industry sectors; however, traditional forecasting methods are unreliable (Bhandari et al., 2023). These methods include applying math formulas, information, and complex models to ocean data such as temperature and then analyzing how the ocean and the weather change in unison (Brunke et al., 2023). The problem lies within the wavering accuracy of these traditional forecasting techniques (Chahari et al., 2022). Without a more reliable forecasting technique, unforeseen weather phenomena continually threaten these aspects of society (Cutler et al., 2022).

Purpose of the Study

To further research linear regression, a potentially revolutionary forecasting method, I will build a regression model to predict weather phenomena such as wind, precipitation, and temperature over 10-day periods. The research aims to evaluate the model's effectiveness and accuracy to determine if linear regression is a more robust alternative to traditional forecasting methods. This study will use the engineering design process to answer the question: How accurate are linear regression models for weather prediction? The purpose of the research is to understand the elements of weather phenomena, the limitations of current forecasting techniques, and the linear regression process in an attempt to provide a robust solution to the problem of unreliable weather prediction. This study will utilize an evaluation of quantitative data to contribute to the overall understanding of machine learning and artificial intelligence, as well as to improve the knowledge of forecasting methods and the reliability of weather reports. The end goal of this study is to create a web application that allows users to create their own weather predictions up to a year in advance.

Research Questions

Q1: How do regression models consume large amounts of data, train themselves, and then make predictions based on that data?

Q2: How accurately can linear regression models predict weather elements such as rain, temperature, and wind?

Hypotheses

H1**:** Using a linear regression model for weather prediction is more accurate over longer time periods.

H2**:** A linear regression model is more accurate with increased training data.

Significance of the Study

This study will benefit all fields reliant on accurate weather reports: agriculture, energy production, travel, event planning, and public safety. Weather forecasts are crucial to the economic and social stability of the world (Cutler et al., 2022; Jaseena & Kovoor, 2022; Kumar & Nabi, 2019). Therefore, a more accurate forecasting method would eliminate unpreparedness for severe weather conditions, such as sudden freezes, storms, and extreme temperatures (Cutler et al., 2022; Fink & Lemburg, 2022). A more reliable forecasting technique would also increase overall economic efficiency since businesses, farmers, and civilians would no longer have to deal with weather uncertainty (Cutler et al., 2022). For example, individuals could travel and plan events further into the future without concern about possible precipitation and varying temperatures (Bhandari et al., 2023). Lastly, this study progresses the application of artificial intelligence and machine learning to new fields, opening new opportunities for future innovations and improving the standard of living. Through its guiding research questions, it will contribute to the overall understanding of machine learning, weather, and artificial intelligence.

Definition of Key Terms

Linear regression

Linear regression models are machine learning algorithms that consider a wide range of training data to analyze patterns and identify relationships between one or more independent variables and the dependent variable (Chahari et al., 2022).

Forecasting

Weather forecasting is the process of estimating the state of the atmosphere in a specific location at a future time (Baral et al., 2020; Holmstrom et al., 2016).

Summary

In summary, weather forecasts are crucial to the economic and social stability of the world (Cutler et al., 2022; Jaseena & Kovoor, 2022; Kumar & Nabi, 2019). Current forecasting, however, lacks accuracy and creates inefficiencies in many sectors of industry, such as power generation and agriculture (Cutler et al., 2022). To solve this issue, researchers are looking to linear regression, which uses artificial intelligence and machine learning algorithms to predict atmospheric phenomena based on vast amounts of training data (Kumar et al., 2017). My study not only examines the efficacy of linear regression but also the issues with current forecasts, the importance of weather forecasting, and the process of forecasting. It aims to create and evaluate the linear regression model to determine if it is an improved alternative to traditional forecasting methods. If the linear regression model proved to be more reliable, it would benefit weather-related industries and decrease unpreparedness among civilians (Cutler et al., 2022; Fink & Lemburg, 2022). This project also progresses the application of artificial intelligence and understanding of machine learning to improve the human standard of living.

Chapter 2: Literature Review

Weather forecasting is the process of estimating the state of the atmosphere in a specific location at a future time (Baral et al., 2020; Holmstrom et al., 2016). Individuals have used this concept since ancient history, and it relies on pattern identification and event recognition (Kumar et al., 2017). Now, as the world has evolved, so have weather forecasting techniques, but there is still vast room for improvement (Holmstrom et al., 2016). According to the National Oceanic and Atmospheric Administration (NOAA, 2024), a seven-day forecast is only correct about 80%

of the time, and for a 10-day forecast, this accuracy falls to 50%. These unreliable forecasts can have significant implications for today's society in various sectors of industry and public safety (Cutler et al., 2022; Kumar & Nabi, 2019). Therefore, understanding the process of weather prediction and weather causation is crucial to create potential solutions for modern forecasting techniques. One of these promising solutions, linear regression, involves artificial intelligence and machine learning algorithms to predict atmospheric phenomena based on vast amounts of training data (Kumar et al., 2017). This review will analyze the process of linear regression, the aspects of current forecasting, and the modern opinions of linear regression to explore possible implementations of this new forecasting method.

Linear Regression Using Machine Learning

As stated by Chahari et al. (2022), linear regression models are machine learning algorithms that consider a wide range of training data to analyze patterns and identify relationships between one or more independent variables and the dependent variable. The model quantifies how the dependent variable changes as the independent variables assume different values (Kumar et al., 2017). This is modeled by a function given by the general equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

where y is the dependent variable, x_1, x_2, and x_n are independent variables, β_0, β_1, β_2, and β_n are parameters/coefficients, and ϵ is error (Kumar et al., 2017). Machine learning algorithms assign values to the parameters and then continue reassigning values until they minimize the errors between the function and training data (Baral et al., 2020). Once it produces a function with minimized error, it can then use the function to predict the dependent variables based on new inputs (Bhandari et al., 2023). These models do not require knowledge of the physical processes that control the weather; Instead, they make predictions based solely on the correlations observed between variables (Jaseena & Kovoor, 2022). This pattern identification ability makes regression models robust against the irregularity of the atmosphere and a viable alternative for current techniques (Baral et al., 2020; Holmstrom et al., 2016; National Centre for Atmospheric Science [NCAS], 2023).

Evolution of Weather Forecasting

Weather forecasting originated in ancient times when humans began to understand patterns and seasonal changes in the atmosphere (Kumar et al., 2017). In this era, humans applied their knowledge of folklore, animal behavior, and plant life cycles to sense when weather was going to be severe (Kumar & Nabi, 2019). In the 17th century, Aristotle criticized the use of theories for weather prediction and stated that accurate forecasting requires more reliable measuring instruments (Kumar & Nabi, 2019). This suggestion led to a series of innovations, such as the thermometer by Galileo in 1643, the anemometer by Hooke in 1667, and later weather balloons and radars in the 1940s (Kumar & Nabi, 2019). These inventions spurred forecasting to numerical weather prediction, which is where it is currently (Cutler et al., 2022). Despite modern technology, Caselles et al. (2022) state that scientists continuously improve these numerical models, suggesting the possibility of perfect forecasting in the future. Many signs point to artificial intelligence and machine learning (Kumar et al., 2017).

Current Forecasting Technique

As stated earlier, modern forecasting uses numerical weather prediction, which begins with data collection (NCAS, 2023). This data collection step involves gathering information regarding the state of the atmosphere, such as pressure, humidity, precipitation, wind, temperature, and satellite images (NCAS, 2023; NOAA, 2024). Then, scientists input this data into computer models, mathematical formulas, or simple equations to simulate how the weather will change over time (Kumar & Nabi, 2019; NCAS, 2023). According to Baral et al. (2020), scientists mostly use physical simulations computed through thermodynamics and fluid dynamics equations to track changes in the atmosphere. However, despite their complexity, they cannot withstand the irregularity of the atmosphere (Baral et al., 2020). This inability creates limitations in current forecasting methods, establishing opportunities for alternative methods such as linear regression (Baral et al., 2020; Bhandari et al., 2023; Holmstrom et al., 2016).

Limitations of Current Forecasting

Numerous studies have analyzed the use of linear regression models for weather prediction, and many aimed at addressing these limitations of current forecasting methods. For example, traditional techniques have limited efficiency, as Bhandari et al. (2023) stated that traditional techniques require large amounts of information and computational resources. To address this, Bhandari et al. (2023) analyzed both the efficiency and resource management of weather regression models. On the contrary, Kumar et al. (2017) found that traditional methods have limited output possibilities, thus creating their study of linear regression for rainfall yield because current data mining prediction methods did not produce an estimate of the rainfall. A more commonly mentioned limitation of current forecasting techniques other than accuracy is the time frame (Baral et al., 2020; Fink & Lemburg, 2022; Holmstrom et al., 2016; NOAA, 2024). According to Baral et al. (2020) and Holmstrom et al. (2016), due to a lack of understanding of atmospheric processes, some natural or physical factors limit the accuracy of weather forecasts to a 10-day period; however, the robust nature of regression models enables researchers or scientists to address this issue. Many researchers agree that linear regression models can solve these limitations in current forecasting models, but there are societal impacts researchers must first focus on addressing (Baral et al., 2020; Bhandari et al., 2023; Holmstrom et al., 2016).

Societal Impacts of Weather Reports

As previously stated, weather forecasting techniques are imperfect and prone to mistakes (Bhandari et al., 2023). However, to best minimize these errors in future forecasting methods, one must identify how unreliable forecasts affect various areas of society (Cutler et al., 2022). Accordingly, Chahari et al. (2022) identified the main weather-affected areas as travel, farming, mining, and energy production, which Cutler et al. (2022) supported through a study on the economic impact of weather forecasts on work commutes, wind energy, and agriculture. The study concluded that misinformation regarding precipitation, wind speed, and temperature causes inefficiencies among businesses and workers, incurs unnecessary costs for renewable energy companies, and significantly decreases agricultural output and revenue (Cutler et al., 2022). While these studies focus on the economic impacts, they ignore the dangerous risks associated with inaccurate forecasts. For example, in August of 2003, more than 700,000 excess deaths occurred due to the unpredictability of a heatwave, which brought drought and dryness to Europe (Fink & Lemburg, 2022). Other studies, such as one done by Jaseena and Kovoor (2022), also emphasize the importance of safety when predicting weather, stating that accuracy is necessary to protect property and people from floods, storms, and tornadoes. Acknowledging these affected sectors helps researchers employ forecasting solutions that generate the right results and consider the correct variables (Jaseena & Kovoor, 2022). For instance, to eliminate the unnecessary costs among wind farm corporations, a solution must aim to increase the accuracy of wind prediction and consider variables attributed to wind, such as air pressure (NCAS, 2023). Improving the predictability of precipitation or evaluating cloud coverage would be inefficient for this issue

(Cutler et al., 2022). Therefore, researchers must identify the weather-affected industries in order to create a solution for more reliable forecasting (Cutler et al., 2022).

Analysis of Past Regression Models

After identifying the concerns of current forecasting techniques, researchers employed their models and analyzed the results. Chahari et al. (2022) studied four machine-learning models, Ridge, Random Forest, Linear Regression, and Decision Tree, and utilized them to predict the temperature of 1000 data points. Although the linear regression models produced reasonable accuracy, it was the lowest among the other machine learning models. Similarly, Baral et al. (2020) focused their model solely on predicting the maximum and minimum temperature of the following day. They found a root mean square error of 1.6 ºC and 1.3 ºC, respectively (Baral et al., 2020). This is low compared to other studies, suggesting that their model is robust for temperature forecasting in that region (Baral et al., 2020).

Despite these findings, the previous studies do not compare the regression models to current forecasting techniques. Kumar et al. (2017) determined that their model produced rainfall yield predictions more accurately than current methods. On the other hand, both Holmstrom et al. (2016) and Bhandari et al. (2023) disagreed and claimed that traditional methods are usually more accurate than those of their linear regression models. The discrepancies most likely come from the type of study conducted as Holmstrom et al. (2016) developed their model to estimate temperature over a period of seven days, and Bhandari et al. (2023) tested multiple weather conditions. Although they had disappointing accuracy results, they both agreed that linear regression models are more efficient and could produce more promising results than current methods over a more extended time period (Bhandari et al., 2023; Holmstrom et al., 2016). This opens possibilities for future studies.

Summary

In summary, numerous areas of the economy and public safety, such as agriculture, power production, and travel, are dependent on the reliability of weather reports (Caselles et al., 2022; Cutler et al., 2022; Jaseena & Kovoor, 2022; Kumar & Nabi, 2019). Unfortunately, current forecasting methods have limitations in accuracy, time, and computational requirements, threatening these aspects of society (Baral et al., 2020; Bhandari et al., 2023; Fink & Lemburg, 2022; Holmstrom et al., 2016; NOAA, 2024). Due to its efficiency and resource management, the linear regression model has recently shown promise as a viable alternative to current weather forecasting solutions (Bhandari et al., 2023; Kumar et al., 2017). Many studies that evaluated linear regression stated that the accuracy is underwhelming (Bhandari et al., 2023; Chahari et al., 2022; Holmstrom et al., 2016). However, according to Holmstrom et al. (2016), these models may be more accurate with increased data collection and better suited for extended periods of time. This assertion opens avenues for future studies to develop a model on a wider range of variables and test the prediction accuracy further into the future. An ideal study would create a regression model and compare it to the 50 percent accuracy rate of current models for a 10-day forecast in an attempt to solve all problems caused by unreliable forecasts (NOAA, 2024). It would increase the time period of the test and include the suggestion made by Holmstrom et al. (2016) for more extensive data collection. Therefore, it builds on previous studies and includes elements not formerly researched to further attempt to solve the problem of unreliable forecasting.

Chapter 3: Research Method

This project explored the accuracy and effectiveness of linear regression, a machine-learning algorithm utilized for weather prediction. Currently, many sectors such as power production, agriculture, and public safety rely on the accuracy of weather reports to maintain business, prepare for hazards, and plan for the future (Caselles et al., 2022; Cutler et al., 2022; Jaseena & Kovoor, 2022; Kumar & Nabi, 2019). However, current forecasting methods have limitations in accuracy, time, and computational requirements, threatening these vital aspects of society (Baral et al., 2020; Bhandari et al., 2023; Fink & Lemburg, 2022; Holmstrom et al., 2016; NOAA, 2024). To combat this, many researchers have been looking to the innovation of machine learning and artificial intelligence as a viable component of future, more accurate forecasting (Bhandari et al., 2023; Kumar et al., 2017).

My engineering study aimed to develop and analyze the accuracy of a linear regression model utilized for weather forecasting. The research aims to evaluate the model's effectiveness and accuracy to determine whether linear regression is a more robust alternative to traditional forecasting methods. The underlying hypotheses for this research are as follows: Using a linear regression model for weather prediction is more accurate over longer time periods (H1), and a linear regression model becomes more accurate as training data increases (H2).

To evaluate these assumptions, the model underwent a series of tests and alterations until it had minimal marginal error between the validation data and actual data. I then used the model to weigh weather predictions against actual test data. This analytical evaluation, along with the development of the model, can provide weather-affected industries with a solution or at least an advancement in the problem of forecast reliability.

Research Design and Method

To thoroughly analyze the accuracy of weather reports with linear regression, one must build a model (Holmstrom et al., 2016). I utilized various software packages along with my Python coding expertise to pursue this feat and generate a model. I used the engineering design process, quantitative data collection, and error analysis to build a model and analyze its effectiveness. As previously stated, linear regression models first need a training set of data with the desired outputs and inputs from which it can self-learn (Holmstrom et al., 2016). My dataset came from the National Centers for Environmental Information (NCEI, 2024), which produces daily summaries composed of weather elements such as precipitation, temperature, and wind speed. My specific dataset is the daily summaries from the Atlanta Hartsfield-Jackson Airport from the dates 1/1/1990 to 3/3/2024. Using this dataset combined with a Python package, Scikit-learn, I created a linear regression model to predict the weather over 10-day periods. To achieve the best accuracy, I ran the model and continuously re-adjusted it in terms of alpha value, input variables, and size of the dataset to pinpoint the factors that produced outcomes that most closely resembled the desired outputs. I chose this process specifically for its robust design and thorough revising, which yielded a model with the greatest accuracy. This study utilized quantitative and marginal error analyses to create a linear regression model with minimized error and then used the model to predict outcomes.

In order to address the goal of the study and analyze the accuracy of the regression model, I used the model to predict 10-day weather forecasts for certain weather elements, such as average temperature. I then compared these reports with the test outputs provided by the NCEI (2024) to evaluate the accuracy using marginal error analysis. I tested the model's validity through criterion validity since it is feasible and typically regarded as the best measure (Middleton, 2023). By utilizing cross-examination, I can treat the accuracy found in the validation stage as criteria and compare it with the accuracies produced in the test stage. These should yield similar results for high validity (Middleton, 2023). To test the reliability, I utilized the test-retest method, which ensures consistency among results (Middleton, 2023). Since I am producing many 10-day forecasts, the accuracies should be similar to the validation set over time as the test re-test method states, ensuring the accuracy of results. My goal was to use the model and generate 32 10-day forecasts of daily average temperature with their respective accuracies to evaluate overall mean accuracy and distribution. Finally, I aimed to implement the model into a web application that allows users to generate their own forecasts. This will accomplish the study's aim to evaluate the effectiveness of linear regression models for weather prediction as well as generate a regression model for weather prediction. I designed this research process to produce the most accurate regression model and yield the best predictions for overall analysis.

This project also contains assumptions and limitations that need discussion. As previously stated, I do not manually collect the data. Instead, I requested it through the NCEI (2024) search service, and therefore, I must assume they collect the data with accuracy and precision. According to their website, the data comes from weather/climate stations, airplanes, radars, weather balloons, moored buoys, ships, satellites, and computer models across the globe, and as a government agency, their data collection should be foolproof (NCEI, n.d.). If there were issues within the agency's data collection method, it would affect the internal validity of my study as it would generate issues within the development of my data and model, not the generalization of results to the situational setting (Andrade, 2018). Another assumption made is that the software and code I am manipulating produce no implications on the accuracy of the model. Fortunately, Scikit-learn (Sklearn) is Python's most robust and useful machine-learning package; therefore, it should not be a problem. Otherwise, it would affect the internal validity of the model (Biswal, 2023). A limitation within the design itself lies within the software chosen. There may be a better available machine learning packages with better accuracy and computational power in the future. This would affect the external validity as linear regression may evolve in the future, which has the potential to change the findings and generalizations of this study in the future. Overall, the method for this research is very meticulous, and it considers all external factors.

Situational Setting

The situational setting for my linear regression model was a 10 km radius centered at the Atlanta Hartsfield-Jackson Airport Weather Station at 33.62972° latitude and -84.44224° longitude, falling into Clayton and Fulton County (NCEI, 2024). I selected this setting for two reasons: according to the NCEI (2024), this weather station has 100% coverage from the years 1990 to 2024, so it has all its necessary data for model implementation, and its coverage area is a radius of about 10 km, which corresponds to the chosen situational area. I chose the data population of daily summaries consisting of temperature, wind speed, and precipitation readings from 1/1/1990 to 3/3/2024 solely for its vast size. The larger size of the data population, combined with a greater number of input variable values, produces a model with the greatest accuracy (Holmstrom et al., 2016). Since I only evaluated the accuracy of the regression model's predictions against the actual data for this weather station, there is no need to significantly focus on the coverage area (Baral et al., 2020). In other words, all the training, validation, and test datasets come from the same single point, for which the model also generates estimates for. Thus, the coverage area does not affect the model's accuracy (Baral et al., 2020). The coverage area is only important regarding the regression model's physical implementation for forecasting use in a particular location (Bhandari et al., 2023). In addition to this, it is essential to note the chaotic nature of weather and its high variability ([NCAS], 2023). This produces external validity problems when it comes to generalizing weather conditions for different situational locations than the one for which I built the model for (Biswal, 2023). Therefore, due to the selection of the situational setting, the model is only functional for forecasting at the Hartsfield-Jackson Weather Station with its current dataset. However, the analysis of the model's accuracy and its corresponding findings can extend to larger forums because, as stated earlier, the location is irrelevant to its accuracy, and the goal is the evaluation of accuracy (Bhandari et al., 2023). If the one used the model for physical implementation, they should also use a dataset for that specified location. As noted previously, I assumed the data provided by the NCEI is precise, as they made a statement about how they use radars, weather balloons, weather stations, weather buoys, and additional quality control to ensure collection accuracy (NCEI, n.d.). I address the specific process of their data collection in a later section.

For this study, there is only one situational setting, excluding the rest and creating a delimitation. This study only has data from one location, the Hartsfield-Jackson Airport, and thus, it has one model. The specificity of these decisions allows me to comprehensively analyze and focus on creating the most accurate regression model. Creating many models with different datasets would be very time-consuming, and it is harder to generalize the results to the chosen situational setting (Baral et al., 2020). In addition to this, other locations do not have as much coverage, meaning that much of their information is missing (NCEI, 2024).

Materials and Instruments

This engineering process utilized various materials in addition to the instruments used by the NOAA for the original collection of data. The archived data requested from the NCEI (2024) is easily accessible to anyone with access to an email for various locations. However, it is essential to note the agency's methodology to collect the original data to ensure my study's validity. According to their website, they measure readings by their corresponding instruments (National Oceanic and Atmospheric Administration [NOAA], 2024). These can be sensors and thermometers for temperature readings, an ombrometer for the amount of precipitation, and anemometers for wind speed (NOAA, 2024). The most crucial step apart from this collection of measurements is quality control (NOAA, 2024). This process identifies outliers and examines inconsistencies among local regions to locate the source of the inaccurate readings. After meteorologists review raw data for errors, they publish it for scientific or recreational use (NOAA, 2024).

In addition to data access, software was an essential component of this project. I selected software packages such as Python version 3.11.5, Scikit-learn version 1.4.0, Pandas version 2.2.0, Matplotlib version 3.8.2, and NumPy version 1.26.3 for their uses and up-to-date data processing. Python 3.11.5 is the current coding version in which I developed and ran my functions. The Python file contains all the code and operations for the data cleaning, the regression model development, and the resulting graphs. Pandas is a library for data and the most powerful open-source tool (Toomet, 2024). It structures all data as a data frame (table) and allows users to manipulate columns and rows, quickly transform data, and handle missing data (Toomet, 2024). Using NumPy and Pandas together is popular as NumPy is one of the leading packages for numerical analysis and optimization (Toomet, 2024). The combination of these packages allowed me to organize my data, handle missing values in data, and perform the end analysis. The Scikit-learn package is robust machine learning software containing the overall regression model operations instructions (Biswal, 2023). Matplotlib was the final package I used, and it produced all the necessary graphs for a visual representation of results.

In addition to this, I used a computer capable of installing and running all the software, as well as Excel. When you request data from the NOAA search service, it exports the data in the form of a spreadsheet, and from there, it requires Excel to process the data and use it. In terms of computer requirements, any modern computer with access to the internet can run these programs; therefore, there are no strict restrictions (Biswal, 2023).

As stated earlier, I assumed the data to be accurate and precise. Otherwise, it would weaken the internal validity of the study. However, since the process for collecting data is methodical and comes from a government agency, inaccuracies are unlikely (NOAA, 2024). Another assumption I made is that all the software is state of the art and that, therefore, it will produce the best linear regression model with maximum accuracy. If this was incorrect, I could not make generalizations about the accuracy of linear regression models since I am not studying them at the peak level, thus affecting external validity. Fortunately, based on several sources and computer scientists, these software packages are the most popular and powerful in today's society, making my project robust (Biswal, 2023; Toomet, 2024).

The only limitation that results from the selection of materials corresponds to the room for improvement in the future. While the technology I used is currently state of the art, an innovation may come in the future that renders my linear regression model obsolete (Biswal, 2023; Toomet, 2024). If my technology were outdated, it would threaten external validity, making my findings irrelevant. At the time of this study, the technology is up to date, and thus, the findings are valid.

Operational Definition of Variables

Alpha value

Also known as the regularization parameter, this value controls the balance of a linear regression model's complexity and performance (Baždarić et al., 2021). By using different alpha values, one can make a model more fit to the training data or more fit for prediction (Baždarić et al., 2021).

Overfitting

Overfitting occurs when a machine learning model becomes too closely catered to modeling the training data and loses the ability to generalize to new, unseen data (Baždarić et al., 2021).

Testing dataset

A dataset held back from training and provides an unbiased estimate of the model's performance to assess its performance in real-world scenarios (Goodacre & Xu, 2018).

Training dataset

A dataset that researchers use to train the machine learning model. It serves as the foundation for the model's learning of patterns, relationships, and features (Goodacre & Xu, 2018).

Validation dataset

It is another dataset withheld from training that researchers implement to tune the model's parameters and help prevent overfitting in order to produce the best predictions (Goodacre & Xu, 2018).

Data Collection, Processing, and Analysis

Data collection began with locating a dataset using the NCEI climate search tool (NCEI, 2024). After selecting the Atlanta Hartsfield-Jackson Airport Weather Station and the date range 1/1/1990 to 3/3/2024 for its coverage and size, I ordered it to my email in a .csv format. I did this because Python can easily read and reference this data format. Contained within each daily summary are the variables average wind speed, peak wind gust, precipitation amount, snow amount, average temperature, maximum temperature, and minimum temperature for implementation. Once downloaded in Excel as a .csv, I split the data into three sections: training, validation, and testing. Since the end goal was to have 32 10-day forecasts of average temperatures predicted by the model and compare the reports with the actual data, I made the testing data the most recent 300 days of my dataset. The validation dataset is 30 days separated into three 10-day forecasts. This set comes from the thirty days of data before the testing dataset, and it could be small since its purpose is solely to help me judge what combination of parameters, variables, and alpha value produces the most accurate predictions. I made the rest of the data before the most recent 330 days my training dataset to give the model as much information as possible. To clarify, the range from 1/1/1990 to 4/8/2023 was the training set, the range from 4/9/2023 to 5/8/2023 was the validation set, and the data from 5/9/2023 to 3/22/2024 was the test set.

After marking all my data, I handled missing values/readings using the Pandas "fillna()" method. Then, using Scikit-learn, I built a model using the function "from sklearn.linear_model import Ridge" and the associated methods found on their website. I then fed the model the training data and target data through the "reg.fit()" method, which created the model. After creating the original model, I used it to produce outputs to compare my predictions with the validation dataset using the "reg.predict()" method. To evaluate it and determine what parameters and alpha values would yield the best accuracies, I used the "mean_absolute_error" method from "sklearn.metric". I continuously re-adjusted the parameters and alpha value until my model was no longer increasing in accuracy. Lastly, I predicted 32 10-day forecasts of average daily temperature with the nearly foolproof regression model. By collecting the mean error from each 10-day forecast and having a size of 32 (N=32), I created a distribution to find the true absolute mean. Due to the fact that this project exhibited quantitative results of slight variance, the data does not follow a normal distribution, the observations are not independent of each other, and hypothesis testing is not applicable. Since this project is an engineering design and the end goal is to find the overall accuracy of linear regression for 10-day forecasts, an r-squared evaluation and mean absolute error are better evaluation statistics (Baždarić et al., 2021). The r-squared correlation would represent the correlation between my model and the intended results, and the mean absolute error would allow me to analyze how far off my predictions are. These measurements lead to the conclusions of this study, where the goal is to evaluate the model's effectiveness. In addition to this, finding variation statistics such as minimum and maximum error will help me generalize my findings to real-world forecasting (Bhandari et al., 2023). By having a minimum error, meteorologists can understand their worst-case scenario, allowing them to make a conscious decision on whether to implement a linear regression model (Bhandari et al., 2023).

Assumptions

I addressed most of the assumptions throughout this design process earlier, but I will reaffirm them. The first assumption was that the original data source collected it with precision and accuracy. According to their website, the NCEI (n.d.), the methodology has numerous steps to ensure the accuracy and quality control of the weather data. If this was not the case, it would affect internal validity as poor data would compromise the development of the model (Biswal, 2023). However, due to the precautionary steps taken, it is doubtful. Another assumption exists within the software chosen. I also assumed that the software utilized for this project is the best in its area and has no implications for the model's accuracy. If this were not the case, it would affect the external validity of my study and the ability to generalize my results (Biswal, 2023). Moreover, my study would be inaccurate because I would not be testing the full potential of regression models. Fortunately, various sources have confirmed that these software packages are the most advanced today, so this protects external validity (Biswal, 2023; Toomet, 2024). The last assumption I made was that the finalized regression model would be the one with the least amount of error. If this were incorrect, my results would underestimate the true accuracy of the linear regression model and external validity (Biswal, 2023). Luckily, I re-adjusted the parameters and re-ran the model until I reached a maximum, and it started declining. This allowed me to find the set of parameters that truly minimize error. Therefore, despite assumptions made, the methodology of the project is robust.

Limitations

As previously stated, the most significant limitation of this project is implementing this model into areas other than the area I developed it for. I trained the model on data from the Hartsfield-Jackson Weather Station and thus only learned from the patterns and variables present at that station. Due to the chaotic nature of weather and its high variability, it is impossible to use this same model in other locations (National Centre for Atmospheric Science [NCAS], 2023). This affects the external validity and my ability to generalize this model's reports to other locations. However, it does not affect the ability to generalize accurate findings of my model since situational setting does not play a role in this evaluation (Bhandari et al., 2023). An additional limitation I had was access to technology. An additional limitation arises from the selection of materials, more specifically, the room for improvement in the future. While the technology I used is currently state of the art, an innovation may come in the future that renders my linear regression model obsolete (Biswal, 2023; Toomet, 2024). Despite the contemporary nature of the software, it could be better or more powerful, creating a better regression model. If this is the case, it would significantly threaten external validity in the future, and my findings would be irrelevant. Fortunately, the software utilized is the current best option, making my findings valid.

Delimitations

One of the biggest delimitations of this project is the exclusion of certain weather elements, such as humidity and UV index. I intentionally left out these measurements of the model's training because the NCEI does not provide these data points in their reports. While humidity plays a prominent role in temperature readings, the regression model has multiple other variables, making it robust against the exclusion of humidity and UV index (Holmstrom et al., 2016). Another delimitation is the exclusion of other datasets and other models. This study only has data from one location, the Hartsfield-Jackson Airport, and thus, it has one model. The specificity of these decisions allows me to fully analyze and focus on creating the most accurate regression model. Creating a large number of models with different datasets would be very time-consuming, and it would be harder to generalize the results to the chosen situational setting (Baral et al., 2020). Lastly, the model will only predict future temperatures and use all the other variables as input. This is because, from the data provided, the model cannot tell what weather conditions derive from certain parameters such as temperature, wind speed, and precipitation. For example, it cannot determine if the weather is cloudy since it does not know the characteristics/parameters of cloudy conditions. Other than these delimitations, I carried out as others would typically infer.

Ethical Assurances

Due to the procedure and data collection method, I had no required ethical assurances to conduct my study. The collection of data and the entity studied are not vertebra and do not conflict with privacy issues. Therefore, this engineering design project does not fall under any ethical considerations and does not need assurances to protect against the threat of ethics.

Summary

In summary, to answer the following hypothesis stating that linear regression models are more accurate over longer time periods of forecasts (H1) and a linear regression model becomes more accurate as training data increases (H2), a quantitative error analysis was necessary. To establish the validity and reliability of the regression model's results, I utilized the test-retest method in conjunction with criterion validity. To carry out the engineering design of the model, I utilized the weather data provided by the NCEI (2024) and split it into a test, training, and validation set. Then, after a series of trials and parameter re-adjustments, I found a regression model with minimized error. Subsequently, I used this model to produce 32 10-day forecasts of average daily temperature and compared them to the actual data to address the hypotheses. I determined this using an r-squared evaluation, variance statistics, and an analysis of mean absolute error. This has not only allowed me to measure the accuracy of my model but also allowed me to gauge the possibilities of error if I deploy the model into actual use.

Chapter 4: Findings and Results

As previously stated, many sectors, such as power production, agriculture, and public safety, rely on the accuracy of weather reports to maintain business, prepare for hazards, and plan for the future (Caselles et al., 2022; Cutler et al., 2022; Jaseena & Kovoor, 2022; Kumar & Nabi, 2019). However, current forecasting methods have limitations in accuracy, time, and computational requirements, threatening these vital aspects of society (Baral et al., 2020; Bhandari et al., 2023; Fink & Lemburg, 2022; Holmstrom et al., 2016; NOAA, 2024). This study aimed at creating a linear regression model to combat these issues and allow users to eventually create their own weather forecasts based on their desired weather parameters.

In order to complete this, the methodology includes the development of a regression model and evaluation of its performance. After training the model on the Hartsfield-Jackson weather data and minimizing its error, I used it to predict average temperature over 32 10-day periods. To answer the hypothesis centered around the model's accuracy, I derived a mean absolute error from each 10-day sample. I found the true mean absolute error for the data population along with various statistical measures, which indicate the spread of error. After evaluation and statistical analysis of the model, I created a website interface to allow users to use this model to forecast their own weather periods. This process utilized the engineering design process along with quantitative data collection to evaluate a linear regression model and build a web application.

Results

After the regression model development, I took a sample of absolute error for 32 10-day periods to analyze the model's accuracy. The assertion of the hypotheses for this study lies in the model's production and how to make the model the most accurate. Due to the nature of the engineering design process, most of the research questions are answered within the research itself. As the end goal is to evaluate its accuracy, one question applies to the collection of data.

The collection of data answered the following question: How accurately can linear regression models predict weather elements such as rain, temperature, and wind? (Q3). With the model developed, it was essential to analyze how well it would predict user forecasts when I constructed the user interface. This would allow users to understand the possible error when they use my web application to predict their desired weather period. After training and validation, I used the datasets from dates 5/9/2023 to 3/3/2024 sectioned into 10-day periods to calculate the absolute mean error of predicting average temperature for each period. The data table example for each date range is shown below.

Table 1

Sample 10-day Forecast

Date	Actual (°F)	Predicted (°F)	Marginal Error
05/09/2023	75	74.366122	0.633878
05/10/2023	74	74.627037	-0.627037
05/11/2023	71	73.284629	-2.284629
05/12/2023	73	71.077609	1.922391
05/13/2023	79	73.776604	5.223396
05/14/2023	77	78.166498	-1.166498
05/15/2023	76	76.287312	-0.287312
05/16/2023	71	75.425858	-4.425858
05/17/2023	69	70.903059	-1.903059
05/18/2023	66	67.667774	-1.667774
Total Error			20.141832
Mean Abs Error			2.014183

Note. N=32 (n=10 for each forecast sample). This is an example sample forecast in the 32-sample set. Total error is the sum of absolute marginal per day, and mean absolute error is the total error divided by the sample size.

This is an example forecast with the date, the actual average temperature for that day, the prediction, and the marginal error. The total absolute error and the mean absolute error are also listed for the forecast range. I took the mean absolute error from each of these tables to find the actual absolute mean error over 32 datasets using my regression model. After collection of data, the 32 datasets had a true mean error of 3.258690 °F, a median of 2.921242 °F, a range of 8.188715 °F, a standard deviation of 1.679242 °F, and a variance of 2.819855 °F. There was an outlier in the original dataset of 9.402226 °F; so after it was removed, the data had 31 values with different statistical measures. The r-squared coefficient across the dataset and the regression model was 0.902931, meaning that the model explains about 90.2931% of the variability observed in the temperatures of 10-day forecasts. In other words, the model captures 90.2931% of the fluctuations in the dependent variable (temperature) around its mean using the predictors (average wind speed, peak wind gust, precipitation amount, snow amount, maximum temperature, and minimum temperature).

Table 2

Stem and Leaf Distribution for 32-sample forecast

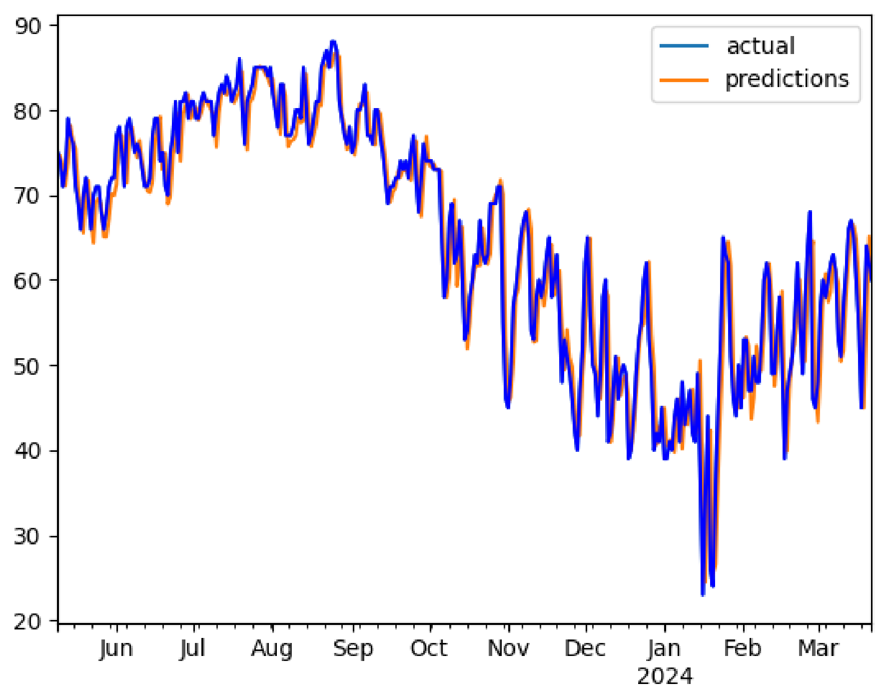
Stem (First Whole Digit)	Leaf (First Decimal Following Whole Digit)
1	2 4 6 6 7 7 7 9
2	0 1 2 3 3 3 4 8 9
3	0 0 6 7 8
4	0 0 4 5 5 5 9
5	1
6	2

Note. Each error for a sample forecast is split into the first whole number and decimal to be viewed as a distribution. Key: 1 | 2 = 1.2 absolute error.

Due to the nature of the data collection and the aim of the study, assumptions for this data collection fall short of qualifications for statistical testing. This project exhibited quantitative results of slight variance, the data does not follow a normal distribution, and the observations are not independent of each other; hypothesis testing and creating a confidence interval are not applicable. However, since the goal of the study is to generate a web application, the measures captured in this data collection are viable for its purpose. Due to the fact that this project is an engineering design, and the end goal is to find the overall accuracy of linear regression for 10-day forecasts and make a web app, an r-squared evaluation and mean absolute error are better evaluation statistics. The graph below captures an overview of the error between the actual average temperature and the model’s predictions for average temperature, where the blue line is the actual and the orange line is the predictions. The vertical axis is the temperature in degrees Fahrenheit, while the horizontal axis is the date.

Figure 1

A Line Graph Showing the Difference Between Actual Temperatures and Prediction Temperatures Over the Test Period



Note. The line graph captures the differences in actual average temperatures and prediction temperatures for the time span 5/9/2023 to 3/22/2023. The blue line is the actual temperature, and the orange is the prediction. The y-axis is the temperature in degrees Fahrenheit, while the x-axis is the date.

Evaluation of Findings

As previously stated, I tested the model's validity through criterion validity since it is feasible and typically regarded as the best measure (Middleton, 2023). By utilizing cross-examination, I treated the accuracy found in the validation stage as a criterion and compared it with the accuracies produced in the test stage. These yielded similar results indicating high validity (Middleton, 2023). To test the reliability, I utilized the test-retest method, which ensures consistency among results (Middleton, 2023). Since I am producing many 10-day forecasts, the accuracies should be similar to each other over time as the test re-test method states, ensuring the accuracy of results. This was true for my data, apart from one outlier indicating that my results are reliable. Based on the reliability and validation results, my data is considered valid.

In addition to this, Baral et al. (2020) focused their model solely on predicting the maximum and minimum temperature of the following day, and they found a mean error of 1.6 ºC and 1.3 ºC, respectively (Baral et al., 2020). It would be reasonable to conclude that as the forecasting period increases, accuracy decreases (NOAA, 2024). Therefore, based on this literature and my extended forecasting period, the true mean would be increased. Since my mean error was 3.060511 °F, it falls within the expected range. The difference between Fahrenheit and Celsius has no implications on this data as mean error solely regards distance from the actual temperature in their respective measurement types. Therefore, the results I received in this study are of high quality.

Summary

In summary, the results of my 32-sample study with the outlier removed yielded a mean error of 3.060511°F, a median of 2.875193 °F, a range of 5.046103 °F, a standard deviation of 1.286090 °F, and a variance of 1.654026 °F. The maximum error of the altered dataset produced by the regression model was 6.259613 °F, and the minimum error was 1.213510 °F. The r-squared coefficient across the dataset and the regression model was 0.902931, meaning that the model explains about 90.2931% of the variability observed in the temperatures of 10-day forecasts. In other words, the model captures 90.2931% of the fluctuations in the dependent variable (temperature) around its mean using the predictors (average wind speed, peak wind gust, precipitation amount, snow amount, maximum temperature, and minimum temperature).

Due to the nature of the study and the aim to create a web application, the statistical measures taken and the procedures chosen are applicable to my study. The data itself is viable and accurate due to test-retest reliability and criterion validity methods, ensuring consistency among data points. In addition to this, a study done by Baral et al. (2020) found errors of 1.6 °C and 1.3 °C. It is expected that for more extended periods of forecast, accuracy will be decreased (NOAA, 2024). This explains my greater variation of predictions but still allows me to draw meaningful conclusions.

Chapter 5: Implications, Recommendations, and Conclusions

To reiterate, current forecasting methods have inabilities and limitations in accuracy, time, and computational requirements, threatening vital aspects of society (Baral et al., 2020; Bhandari et al., 2023; Fink & Lemburg, 2022; Holmstrom et al., 2016; NOAA, 2024). Sectors such as power production, agriculture, and public safety rely on the accuracy of weather reports to maintain business, prepare for hazards, and plan for the future, so society needs a more reliable forecasting system (Caselles et al., 2022; Cutler et al., 2022; Jaseena & Kovoor, 2022; Kumar & Nabi, 2019). This study developed a regression model, analyzed its error, and created a web application to allow users to forecast their weather based on their specified inputs and forecast periods. Using 32 10-day forecast samples and a massive amount of weather data, I was able to evaluate the regression model to help users understand the risks associated with using this method. In predicting average temperature across 31 10-day forecasts with the outlier removed, the sample produced a mean error of 3.060511°F, a median of 2.875193 °F, a range of 5.046103

°F, a standard deviation of 1.286090 °F, and a variance of 1.654026 °F. The maximum error of the dataset was 6.259613 °F, and the minimum error was 1.213510 °F. These findings allowed the creation of a web application that helps users plan future activities further out than typical forecasts with the corresponding possible error.

Implications

This study answers the following research questions: How do regression models consume large amounts of data, train themselves, and then make predictions based on that data? (Q1). As well as how accurately can linear regression models predict weather elements such as rain, temperature, and wind? (Q2). Due to the nature of the engineering design process, the research and creation of the model answers one question, while the data collection process and results answer the other question. This study also utilized the hypotheses in the development stage of the web application as they helped build the regression model with the greatest accuracy for its implementation in web development.

Research played a heavy role in determining what industries are impacted by weather forecasts as well as how to build a linear regression model for weather prediction. Therefore, the question of how regression models consume data, train themselves, and then make predictions was an essential aspect of my study (Q1). To answer this, I looked at how I created and refined my regression model as well as how previous researchers constructed theirs. As stated by Chahari et al. (2022), linear regression models are machine learning algorithms that consider a wide range of training data to analyze patterns and identify relationships between one or more independent variables and the dependent variable. The model quantifies how the dependent variable changes as the independent variables assume different values (Kumar et al., 2017). This is modeled by a function given by the general equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

where y is the dependent variable, x_1, x_2, and x_n are independent variables, β_0, β_1, β_2, and β_n are parameters/coefficients, and ϵ is error (Kumar et al., 2017). Once it produces a function that matches the training data, it can then use the function to predict the dependent variables based on new inputs (Bhandari et al., 2023). My study utilized an application called Scikit-learn to provide machine learning software for the regression model. This software introduced an additional parameter labeled the alpha value, which controlled how the model would generate predictions that either closely resembled the training data or used more inference/generalization. A small alpha value (close to zero) allows the model to fit the training data more closely, while a high alpha value (close to one) allows the model to generalize more (Baždarić et al., 2021). In this study, the optimal alpha value was α=0.10, which made the model generalize less because the prediction dates were not extensively far from the training dates. In other words, greater generalization was not needed since each sample prediction forecast came right after the training data, and they were not significantly far apart in time. Thus, this explains how machine learning regression models take in data, train themselves, and predict results.

The following research question for this study was answered using the data collected: How accurately can linear regression models predict weather elements such as rain, temperature, and wind? (Q2). By taking the true mean error gathered from the sample, we know the average absolute error for the model is roughly low and about 3.060511°F. However, because the distribution of our samples is right-skewed, the median is a more reliable measure of center, producing an error of 2.875193 °F. The maximum error of the dataset was 6.259613 °F, and the minimum error was 1.213510 °F. These numbers are significantly important for users to understand their chances of receiving wrong reports. Probably the most significant of them all is the r-squared correlation of 0.902931. This indicates that the regression model is 90.2931% more accurate at predicting outputs than it would be using the average output as the prediction. This accuracy also explains how accurate the model would predict other variables due to the nature of machine learning regression. It can take the same training data from the previous study and forecast a different variable using the model it has already constructed (Baral et al., 2020). Therefore, this notes the accuracy of the regression model in predicting weather elements such as temperature, wind, and precipitation, thus answering the research question.

The model development in this study utilized the underlying hypotheses to create a model with minimized error. Although they were not statistically evaluated, they played a role in developing the regression model as the purpose of hypotheses in the engineering design process. The hypotheses were as follows: Using a linear regression model for weather prediction is more accurate over longer time periods (H1), and a linear regression model is more accurate with increased training data (H2). Based on the analysis of my validation data and preliminary research, H1 was unsupported, but H2 held true. These were the original parameters for this study because, according to Holmstrom et al. (2016), regression models may be more accurate with increased data collection and better suited for extended periods of time, so I used and evaluated this recommendation. From the usage of the validation set and altering the range of the training data, I could see that the error of my model steadily decreased as my training dataset became larger. This led to the decision to choose such a large range of dates from the NOAA climate data archive as my training dataset (NOAA, 2024). While this hypothesis (H2) held true, extended periods of forecasting did not yield more accurate forecasts, as Holmstrom et al. (2016) predicted. There was no noticeable trend in accuracy as I altered the forecast period; therefore, I did not gather information on numerous forecasting periods, only 10-day forecasts. This is acceptable as the main goal of this study was to develop the regression model, and the hypotheses allow me to optimally create one for my web application. Unfortunately, these hypotheses are only applicable to this study and the way I created this regression model. They cannot be supported or disproved since this research does not statistically test these assumptions.

Additionally, the study's statistical measures are limited to the Atlanta Hartsfield-Jackson Airport weather station. Due to the chaotic nature of weather and its high variability, different weather areas can have much different data and, thus, different error results (NCAS, 2023). The error analysis and sample collected in this study only serve as a preliminary measure to caution users about the possibility of error when using this model to cast their own forecasts using the web application. Therefore, it is very important to interpret the results only in the context of this situational setting.

The purpose of this study was to analyze the accuracy of linear regression and develop a web application that allows users to use this robust method of prediction to create their own forecasts. These forecasts can be made further into the future and are more accurate than the ones produced by traditional means, which are only correct about 50% of the time for 10-day forecasts (NOAA, 2024). This study proves that linear regression is a viable alternative, and it can be used to predict 10-day forecasts with some degree of error but 84% accuracy for the Atlanta Hartsfield-Jackson Airport weather station. However, the aim of the study is not focused on this. The true goal of this study was to use the information gathered and the regression model built to create a web application that allows users to input their own weather data from any location and select a desired forecast. Most current methods of forecasting do not even offer larger than 10-day forecasts, while linear regression methods and my web application allow users to forecast up to a year in advance with their possible degree of error (NOAA, 2024). They can simply input their .csv file, select the training parameters, select the output variable, and then receive their predicted weather outputs.

This engineering study and its end goal help combat the issues proposed by previous literature and provide a solution. As previous literature states, current forecasting methods have inabilities and limitations in accuracy, time, and computational requirements; linear regression does not (Baral et al., 2020; Bhandari et al., 2023; Fink & Lemburg, 2022; Holmstrom et al., 2016; NOAA, 2024). Therefore, the web application developed in this study can benefit the societal areas that are subject to the problems of unreliable forecasting. Sectors such as power production, agriculture, and public safety rely on the accuracy of weather reports to maintain business, prepare for hazards, and plan for the future, and my web application makes extensive forecast planning possible (Caselles et al., 2022; Cutler et al., 2022; Jaseena & Kovoor, 2022; Kumar & Nabi, 2019). While the website application and regression model do not completely solve the issue of unreliable forecasting, as there is always some degree of error, it allows users to have more control over the date and period they want to predict. It also increases user confidence in forecasts due to the knowledge of possible errors. If consumers and businesses know the possible errors associated with their weather predictions, they would have more trust in the service, which would decrease public trust issues posed by other literature (Caselles et al., 2022). Therefore, it addresses some of the issues in the field of forecasting and advances the tools available to society for weather prediction.

Recommendations

As stated already, this study is applicable to many aspects of society, mainly in fields affected by unreliable forecasting. In terms of travel, while users are currently limited to 14-day forecasts, the implementation of this model and web application allows users to generate larger forecasts for more realistic travel planning. While these forecasts being further in the future would likely be accompanied by larger possible errors, users already understand the risks and thus are more trusting of the prediction service (Caselles et al., 2022). As my study concluded a median error of 2.875193 °F for a 10-day forecast and about a 90% r-squared correlation, it is reasonable to believe the model is fairly accurate for forecasts of this size. In addition to travel and public trust, businesses can also use this web application. For example, agriculture companies rely on weather forecasts for crop growth, energy companies such as wind farms use weather forecasts to gauge energy output, and construction companies rely on forecasts to schedule jobs (Cutler et al., 2022). A forum that gives users access to accurate long-range weather data, such as this web application, helps increase efficiency and profits among these industries (Cutler et al., 2022; Jaseena & Kovoor, 2022).

While the web application allows users to create forecasts a year in advance, it is recommended that they stay as precise as possible to their desired date. In other words, if a trip is planned five weeks in advance, the forecast should be for that date range and no longer. This is due to the fact that, at a certain point, the potential error would be too great for the prediction to mean anything significant. To clarify, a prediction of 70°F with a possible error of 30°F would mean nothing because it could be as hot as 100°F or as cold as 40°F. In the future, I recommend evaluating the trend between possible error and date range to find if there is a method to increase the forecast range of the linear regression model without significantly compromising forecast confidence. This would greatly improve the application's ability as it has more use. To achieve this, I recommend a higher alpha value than should be used to create the regression model. This would allow the model to have more inference but less error (Baždarić et al., 2021). Even more optimally, there should be a way to have the model's alpha value change as the forecasting range increases, but currently, the web application lacks this aspect.

Conclusions

In conclusion, linear regression and the web application created in this project can benefit numerous societal fields such as travel, business, public safety, agriculture, and energy production. By creating a user interface that allows users to forecast periods up to a year in the future, all people can extensively prepare for travel and environmental conditions. Since the model has a median error of 2.875193 and an r-squared value of 0.902931, it can be accurately used to make predictions up to 10-day forecasts and more with the according measure of error.

The study covered the research questions of how regression models consume large amounts of data, train themselves, and then make predictions based on that data. (Q1). As well as how accurately can linear regression models predict weather elements such as rain, temperature, and wind? (Q2). The method regression uses to train itself centers on machine learning as it intakes the parameters, trains itself, and makes a model based on the regression equation and alpha value (Baral et al., 2020). The accuracy of regression models is not perfect; however, it produces a mean error of 3.060511°F, which is helpful for users to prepare for weather without threatening prediction confidence.

I used the underlying hypotheses in this study to build the model. The hypothesis suggesting that a linear regression model for weather prediction is more accurate over longer time periods (H1) was unsupported in my findings, while the hypothesis stating that a linear regression model is more accurate with increased training data (H2) held true.

Although the model is optimal for users and advances current forecasting, it is still not perfect. In the future, to advance this study, I recommend creating another regression model but one with a larger alpha value, which allows the model to generalize more for more future time periods without compromising accuracy. Overall, this study advanced the field of forecasting and opened up new avenues for future research.

References

Andrade C. (2018). Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian journal of psychological medicine*, 40(5), 498–499. https://doi.org/10.4103/IJPSYM.IJPSYM_334_18

Baral, A., Shrestha, S., Shrestha S., Shrestha S., & Tiwari, P. (2020, June 5). Weather prediction using multivariable linear regression. *International Journal of Advanced Social Sciences*, 3(1), 25-31. https://ictaes.org/wp-content/uploads/2020/06/IJASS-Vol_03_No_01/IJASS-V3-N1-4.pdf?ckattempt=1

Baždarić, K., Lucijanic, M., Martinović, A., Salarić, I., & Sverko, D. (2021, September 21). The ABC of linear regression analysis: What every author and editor should know. *European Science Editing*, 2021(47), 1-9. https://doi.org/10.3897/ese.2021.e63780

Bhandari, S., Sairam, K.L., & Sapkota, P. (2023, April). Weather prediction system using linear regression. *International Journal of Research Publication and Reviews,* 4(4), 1219-1222. https://ijrpr.com/uploads/V4ISSUE4/IJRPR11492.pdf

Biswal, A. (2023, April 3). *Sklearn Linear Regression*. Simplilearn. https://www.simplilearn.com/tutorials/scikit-learn-tutorial/sklearn-linear-regression-with-examples

Brunke, M. A., Scheftic, W. D., & Zeng, X. (2023). Seasonal forecasting of precipitation, temperature, and snow mass over the western United States by combining ensemble postprocessing with Empirical ocean-atmosphere teleconnections. *Weather & Forecasting*, *38*(8), 1413–1427. https://doi.org/10.1175/WAF-D-22-0099.1

Caselles, V., Gómez, I., Niclòs, R., Valor, E., Molina, S., & Valor, E. (2022, April 22). Confidence and Communication Preferences on Weather Forecasts among University

Students in Spain. Weather, Climate & Society, 14(3), 737–753. https://doi.org/10.1175/WCAS-D-21-0100.1

Cutler, H., Hartman, B., Hill, R., Hu, Y., Jeon, H., Lu, T., Shields, M., & Turner, D. D. (2022, February). Evaluating the Economic Impacts of Improvements to the High-Resolution Rapid Refresh (HRRR) Numerical Weather Prediction Model. *Bulletin of the American Meteorological Society*, 103(2), E198–E211. https://doi.org/10.1175/BAMS-D-20-0099.1

Fink, A. H., & Lemburg, A. (2022, July 11). Identifying Causes of Short-Range Forecast Errors in Maximum Temperature during Recent Central European Heatwaves Using the ECMWF-IFS Ensemble. *Weather & Forecasting*, 37(10), 1885–1902. https://doi.org/10.1175/WAF-D-22-0033.1

Goodacre, R., & Xu, Y. (2018, October 29). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2(3), 249-262. https://doi.org/10.1007/s41664-018-0068-2

Holmstrom, M., Liu, D., & Vo, C. (2016, December 15). Machine learning applied to weather forecasting. Stanford University. https://cs229.stanford.edu/proj2016/report/HolmstromLiuVo-MachineLearningAppliedToWeatherForecasting-report.pdf

Jaseena, K.U., & Kovoor, B.C. (2022, June 20). Deterministic weather forecasting models based on intelligent predictors: A survey. *Journal of King Saud University - Computer and Information Sciences*, 3(6), 3393–3412. https://doi.org/10.1016/j.jksuci.2020.09.009.

Kumar, D. & Nabi, Z. (2019, January). From equations of physics to ensemble models: The evolution and revolution of weather forecasting. *International Journal of Research and*

Analytical Reviews, 6(1), 505–509. https://www.researchgate.net/publication/360932058_The_Evolution_and_Revolution_of_Weather_Forecasting_A_Review

Kumar, P. N., Prabakaran, S., & Tarun P. S. M. (2017, June). Rainfall prediction using modified linear regression. ARPN Journal of Engineering and Applied Sciences, 12(12), 3715-3718. https://www.arpnjournals.org/jeas/research_papers/rp_2017/jeas_0617_6115.pdf

Middleton, F. (2023, June 22). *Reliability vs. validity in research | Difference, types and examples*. Scribbr. https://www.scribbr.com/methodology/reliability-vs-validity/

National Centre for Atmospheric Science. (2023). *What causes weather?*. Natural Environment Research Council. https://ncas.ac.uk/learn/what-causes-weather/

National Centers for Environmental Information (n.d.). *NCEI map viewer application – Data locator*. *NOAA*. https://www.climate.gov/maps-data/dataset/ncei-map-viewer-application-data-locator

National Centers for Environmental Information (2024). *Climate data online search.* NOAA. https://www.ncei.noaa.gov/cdo-web/search

National Oceanic and Atmospheric Administration. (2024, January 19). *How reliable are weather forecasts?*. NOAA SciJinks. https://scijinks.gov/forecast-reliability/

National Oceanic and Atmospheric Administration. (2024, March 7). *How do weather observations become climate data?*. NOAA. https://www.climate.gov/maps-data/climate-data-primer/how-do-weather-observations-become-climate-data

Toomet, O. (2024, February 28). *Machine learning in python*. University of Washington. https://faculty.washington.edu/otoomet/machinelearning-py/numpy-and-pandas.html

Twain, M. (1887). *English as she is taught. The Century Magazine*, *33*(6), 932–938. https://www.gutenberg.org/files/61212/61212-h/61212-h.htm

Appendix A: Raw Data

Appendix A Table 1

Raw Data Collection of Daily Average Temperature Forecasts: Actual and Predictions

Date	Actual (°F)	Predictions (°F)	Marginal Error
05/09/2023	75	74.36406	0.635937
05/10/2023	74	74.61783	-0.61783
05/11/2023	71	73.28879	-2.28879
05/12/2023	73	71.06185	1.938152
05/13/2023	79	73.73085	5.269152
05/14/2023	77	78.16538	-1.16538
05/15/2023	76	76.28825	-0.28825
05/16/2023	71	75.42019	-4.42019
05/17/2023	69	70.88904	-1.88904
05/18/2023	66	67.69906	-1.69906
05/19/2023	70	65.50271	4.49729
05/20/2023	72	70.06191	1.938092
05/21/2023	69	71.70947	-2.70947
05/22/2023	66	67.05637	-1.05637
05/23/2023	70	64.35569	5.644306
05/24/2023	71	69.15306	1.846942
05/25/2023	71	69.64677	1.353226
05/26/2023	68	68.91467	-0.91467
05/27/2023	66	65.09884	0.901162
05/28/2023	68	65.06825	2.931746
05/29/2023	71	67.2052	3.794798
05/30/2023	72	70.16627	1.833726
05/31/2023	72	69.93972	2.060279
06/01/2023	77	71.24197	5.758028
06/02/2023	78	76.22	1.779995
06/03/2023	75	77.01964	-2.01964
06/04/2023	71	74.09111	-3.09111
06/05/2023	78	71.45849	6.541515
06/06/2023	79	77.19248	1.807521
06/07/2023	77	77.96226	-0.96226
06/08/2023	75	75.90362	-0.90362
06/09/2023	76	74.38443	1.615567
06/10/2023	75	76.4174	-1.4174
06/11/2023	73	73.79254	-0.79254
06/12/2023	71	71.82321	-0.82321
06/13/2023	71	70.48032	0.519682
06/14/2023	72	70.29933	1.700669
06/15/2023	77	71.47509	5.524909
06/16/2023	79	75.87264	3.127356

06/17/2023	79	78.60217	0.39783
06/18/2023	74	79.22311	-5.22311
06/19/2023	75	72.96338	2.036615
06/20/2023	71	74.96262	-3.96262
06/21/2023	70	68.92871	1.071293
06/22/2023	75	69.72771	5.272288
06/23/2023	77	74.71044	2.289555
06/24/2023	81	76.19991	4.800088
06/25/2023	75	79.57297	-4.57297
06/26/2023	81	73.97928	7.020723
06/27/2023	81	79.59195	1.408048
06/28/2023	82	80.57852	1.421478
06/29/2023	79	81.84743	-2.84743
06/30/2023	81	78.96643	2.033574
07/01/2023	81	80.00973	0.990268
07/02/2023	79	80.31692	-1.31692
07/03/2023	79	78.8082	0.1918
07/04/2023	81	79.57057	1.429426
07/05/2023	82	80.87468	1.125321
07/06/2023	81	81.2371	-0.2371
07/07/2023	81	80.47723	0.522775
07/08/2023	81	79.92772	1.072275
07/09/2023	77	79.55192	-2.55192
07/10/2023	80	75.61572	4.384283
07/11/2023	82	79.71833	2.281668
07/12/2023	83	81.97032	1.029676
07/13/2023	82	82.39372	-0.39372
07/14/2023	84	81.71656	2.283442
07/15/2023	83	82.75629	0.243713
07/16/2023	81	81.63022	-0.63022
07/17/2023	82	80.82459	1.175414
07/18/2023	83	81.59366	1.406335
07/19/2023	86	82.16953	3.830467
07/20/2023	81	84.53928	-3.53928
07/21/2023	76	79.34962	-3.34962
07/22/2023	81	75.30128	5.698719
07/23/2023	82	80.78412	1.215877
07/24/2023	83	81.32051	1.679494
07/25/2023	85	82.40348	2.596515
07/26/2023	85	84.87869	0.121312
07/27/2023	85	85.14512	-0.14512
07/28/2023	85	84.92206	0.077937
07/29/2023	85	84.3113	0.6887
07/30/2023	84	83.73965	0.260352
07/31/2023	85	82.87903	2.120965
08/01/2023	82	83.69946	-1.69946

08/02/2023	80	81.1442	-1.1442
08/03/2023	78	79.04989	-1.04989
08/04/2023	83	77.11128	5.888725
08/05/2023	83	82.91252	0.087478
08/06/2023	77	81.69788	-4.69788
08/07/2023	77	75.68263	1.317373
08/08/2023	77	76.28907	0.710935
08/09/2023	78	76.48484	1.515157
08/10/2023	80	76.94789	3.052109
08/11/2023	80	79.11016	0.889844
08/12/2023	79	78.50336	0.496637
08/13/2023	85	78.73703	6.26297
08/14/2023	81	84.23178	-3.23178
08/15/2023	76	80.03719	-4.03719
08/16/2023	77	75.69381	1.306194
08/17/2023	79	76.92996	2.070045
08/18/2023	81	78.31883	2.681174
08/19/2023	81	80.78136	0.218638
08/20/2023	85	80.4976	4.502404
08/21/2023	86	84.9621	1.037903
08/22/2023	87	85.37531	1.62469
08/23/2023	85	85.69535	-0.69535
08/24/2023	88	85.09199	2.908008
08/25/2023	88	86.54742	1.452584
08/26/2023	87	86.61773	0.382266
08/27/2023	81	86.25489	-5.25489
08/28/2023	79	79.44288	-0.44288
08/29/2023	77	78.29216	-1.29216
08/30/2023	76	75.35833	0.641673
08/31/2023	78	75.62307	2.376935
09/01/2023	75	76.1532	-1.1532
09/02/2023	76	74.71115	1.288854
09/03/2023	80	76.66001	3.339994
09/04/2023	80	80.7049	-0.7049
09/05/2023	81	80.49132	0.508679
09/06/2023	83	80.80422	2.195785
09/07/2023	77	82.03716	-5.03716
09/08/2023	77	76.91056	0.089436
09/09/2023	76	76.4534	-0.4534
09/10/2023	80	75.88632	4.113684
09/11/2023	80	79.99425	0.00575
09/12/2023	77	79.54727	-2.54727
09/13/2023	75	76.82703	-1.82703
09/14/2023	72	74.54511	-2.54511
09/15/2023	69	71.1965	-2.1965
09/16/2023	71	69.16682	1.833184

09/17/2023	71	70.42494	0.575062
09/18/2023	72	70.7967	1.2033
09/19/2023	72	72.78893	-0.78893
09/20/2023	74	72.11598	1.88402
09/21/2023	73	73.09515	-0.09515
09/22/2023	74	72.77526	1.224743
09/23/2023	72	73.26855	-1.26855
09/24/2023	75	71.72199	3.278012
09/25/2023	77	75.0172	1.982803
09/26/2023	72	76.35032	-4.35032
09/27/2023	68	70.57785	-2.57785
09/28/2023	72	67.48112	4.518879
09/29/2023	76	72.33134	3.668657
09/30/2023	74	76.89368	-2.89368
10/01/2023	74	73.52176	0.478239
10/02/2023	74	73.28836	0.711636
10/03/2023	73	73.64493	-0.64493
10/04/2023	73	72.93819	0.06181
10/05/2023	73	72.94589	0.054107
10/06/2023	65	72.78587	-7.78587
10/07/2023	58	62.69829	-4.69829
10/08/2023	60	57.96353	2.036471
10/09/2023	67	60.33827	6.66173
10/10/2023	69	66.4778	2.5222
10/11/2023	62	69.46903	-7.46903
10/12/2023	63	59.29339	3.706608
10/13/2023	67	62.33347	4.666529
10/14/2023	61	66.26507	-5.26507
10/15/2023	53	59.67066	-6.67066
10/16/2023	54	51.92683	2.073173
10/17/2023	58	54.82428	3.175723
10/18/2023	60	59.53936	0.460638
10/19/2023	63	61.38484	1.61516
10/20/2023	62	61.71732	0.282681
10/21/2023	67	61.6941	5.305897
10/22/2023	63	66.13156	-3.13156
10/23/2023	62	63.03099	-1.03099
10/24/2023	63	61.89792	1.102078
10/25/2023	69	63.08342	5.916577
10/26/2023	69	69.01398	-0.01398
10/27/2023	69	70.02065	-1.02065
10/28/2023	71	70.3666	0.633396
10/29/2023	71	71.80585	-0.80585
10/30/2023	55	70.04394	-15.0439
10/31/2023	46	53.37583	-7.37583
11/01/2023	45	45.60413	-0.60413

11/02/2023	49	46.17651	2.823495
11/03/2023	57	50.35654	6.643461
11/04/2023	59	57.96978	1.030223
11/05/2023	62	58.90821	3.091791
11/06/2023	65	62.78347	2.216526
11/07/2023	67	65.34518	1.654822
11/08/2023	68	67.7816	0.218402
11/09/2023	65	68.34452	-3.34452
11/10/2023	54	65.90984	-11.9098
11/11/2023	53	52.68581	0.314187
11/12/2023	58	52.8428	5.157198
11/13/2023	60	58.55321	1.446785
11/14/2023	58	59.68584	-1.68584
11/15/2023	60	56.94434	3.055656
11/16/2023	63	59.06532	3.934683
11/17/2023	65	63.58022	1.419785
11/18/2023	58	64.19389	-6.19389
11/19/2023	60	58.21659	1.783405
11/20/2023	63	59.58276	3.417242
11/21/2023	56	61.09147	-5.09147
11/22/2023	48	54.65038	-6.65038
11/23/2023	53	49.43167	3.568326
11/24/2023	51	54.17561	-3.17561
11/25/2023	49	50.8871	-1.8871
11/26/2023	46	49.78029	-3.78029
11/27/2023	42	45.35751	-3.35751
11/28/2023	40	41.75146	-1.75146
11/29/2023	47	41.74153	5.258472
11/30/2023	52	48.95888	3.041115
12/01/2023	62	52.70295	9.297052
12/02/2023	65	62.81156	2.188441
12/03/2023	56	64.93879	-8.93879
12/04/2023	50	55.78406	-5.78406
12/05/2023	49	50.18764	-1.18764
12/06/2023	44	48.56328	-4.56328
12/07/2023	49	46.00963	2.990365
12/08/2023	58	50.66068	7.339324
12/09/2023	60	58.85646	1.143538
12/10/2023	41	58.14155	-17.1416
12/11/2023	42	40.91234	1.087656
12/12/2023	47	43.60391	3.396085
12/13/2023	51	48.95888	2.041115
12/14/2023	46	50.86452	-4.86452
12/15/2023	49	46.43903	2.56097
12/16/2023	50	48.27854	1.721462
12/17/2023	49	48.8148	0.185205

12/18/2023	39	46.39636	-7.39636
12/19/2023	40	39.94087	0.059127
12/20/2023	44	42.24878	1.751217
12/21/2023	49	46.7101	2.289899
12/22/2023	53	51.20812	1.791883
12/23/2023	55	54.51138	0.488618
12/24/2023	60	54.8215	5.178503
12/25/2023	62	57.98738	4.01262
12/26/2023	53	62.15162	-9.15162
12/27/2023	49	53.61359	-4.61359
12/28/2023	40	49.64247	-9.64247
12/29/2023	42	40.65724	1.342756
12/30/2023	41	42.8092	-1.8092
12/31/2023	45	42.3174	2.682605
01/01/2024	39	44.97555	-5.97555
01/02/2024	39	40.50192	-1.50192
01/03/2024	41	40.96922	0.030783
01/04/2024	40	41.33825	-1.33825
01/05/2024	44	39.74978	4.250224
01/06/2024	46	43.90802	2.091981
01/07/2024	41	46.06143	-5.06143
01/08/2024	48	40.14446	7.85554
01/09/2024	43	44.63951	-1.63951
01/10/2024	45	42.99822	2.00178
01/11/2024	47	46.23842	0.761581
01/12/2024	42	47.1149	-5.1149
01/13/2024	41	41.88304	-0.88304
01/14/2024	49	42.084	6.916001
01/15/2024	38	50.52704	-12.527
01/16/2024	23	36.53061	-13.5306
01/17/2024	35	24.45376	10.54624
01/18/2024	44	36.75083	7.249172
01/19/2024	26	42.32049	-16.3205
01/20/2024	24	25.32529	-1.32529
01/21/2024	36	26.58607	9.413929
01/22/2024	45	36.6773	8.322699
01/23/2024	53	45.09206	7.907937
01/24/2024	65	53.38905	11.61095
01/25/2024	63	64.12675	-1.12675
01/26/2024	62	64.50708	-2.50708
01/27/2024	51	61.74704	-10.747
01/28/2024	46	48.65869	-2.65869
01/29/2024	44	44.72708	-0.72708
01/30/2024	50	45.01246	4.987544
01/31/2024	45	49.41259	-4.41259
02/01/2024	53	46.15847	6.841528

02/02/2024	53	53.31136	-0.31136
02/03/2024	47	52.29288	-5.29288
02/04/2024	47	43.67447	3.325533
02/05/2024	51	46.13253	4.867473
02/06/2024	48	52.27276	-4.27276
02/07/2024	48	49.50518	-1.50518
02/08/2024	53	49.35915	3.640848
02/09/2024	60	54.46186	5.538142
02/10/2024	62	60.8337	1.166304
02/11/2024	60	61.92155	-1.92155
02/12/2024	49	57.88473	-8.88473
02/13/2024	49	47.52564	1.474358
02/14/2024	54	50.21328	3.786725
02/15/2024	58	55.04804	2.951956
02/16/2024	50	58.694	-8.694
02/17/2024	39	48.13623	-9.13623
02/18/2024	47	39.8821	7.117901
02/19/2024	49	47.90087	1.099128
02/20/2024	51	51.26082	-0.26082
02/21/2024	56	52.47615	3.523845
02/22/2024	62	55.90723	6.09277
02/23/2024	55	60.06005	-5.06005
02/24/2024	49	53.54895	-4.54895
02/25/2024	58	50.44765	7.552349
02/26/2024	64	57.87297	6.127027
02/27/2024	68	62.71388	5.286116
02/28/2024	46	64.59858	-18.5986
02/29/2024	45	46.40712	-1.40712
03/01/2024	48	43.31179	4.688215
03/02/2024	57	49.57672	7.423281
03/03/2024	60	58.05935	1.940647
03/04/2024	58	60.74302	-2.74302
03/05/2024	60	57.4107	2.589302
03/06/2024	62	58.61184	3.388155
03/07/2024	63	62.51676	0.483239
03/08/2024	61	60.95765	0.042346
03/09/2024	53	59.58447	-6.58447
03/10/2024	51	50.4308	0.569203
03/11/2024	57	51.73274	5.267257
03/12/2024	61	57.70264	3.297359
03/13/2024	66	62.18475	3.815247
03/14/2024	67	66.7052	0.294801
03/15/2024	65	66.41767	-1.41767
03/16/2024	60	64.77615	-4.77615
03/17/2024	54	60.07878	-6.07878
03/18/2024	45	52.52432	-7.52432

03/19/2024	55	44.9663	10.0337
03/20/2024	64	55.1666	8.833396
03/21/2024	62	65.13649	-3.13649
03/22/2024	60	60.72145	-0.72145

Note. This is every daily temperature average value in degrees Fahrenheit from the period 5/9/2023 to 3/22/2024. It contains predictions and actual and marginal errors.