Weather forecasting using Machine Learning

Diaa Salama AbdElminaam[1], Salma Baligh[2],
Hana Alaa [3],Jasmine Hegazy [4], Ahmed Mohamed Said [5]
Faculty of Computer Science, Artificial Intelligence
Misr International University, Cairo, Egypt
diaa.salama[1], salma2005893[2],
hana2000171[3], jasmine2004084[4], ahmed2010587[5]{@miuegypt.edu.eg}

***Abstract*—Predicting the weather is a challenging task for all weather researchers as well as the meteorological service. There have been many different methods for predicting the weather over the years from many years ago. Researchers can now anticipate the weather with greater ease and accuracy because to advancements in science and technology. Modern computer science advancements known as soft computing approaches enable accurate and low-error weather forecasting. Traditional methods for predicting the weather involve using a variety of physics models and historical data. This weather forecast is unstable because of how the weather tends to change. The outlook is uncertain because the weather system has changed..In this paper, we offer various machine learning models that will be trained on historical data and then utilised to forecast the weather with a higher degree of accuracy than the conventional methods.The evaluation of the models based on accuracy reveals that the models work better than expected and can be adopted as an advanced way to predict the weather in a more accurate manner and with less time. [1]. [2].**

Keywords: Weather forecast, Machine Learning ,ARIMA,SARIMA.

I. Introduction

The world's weather is constantly and quickly changing. Today's society relies heavily on accurate forecasts. We significantly rely on weather forecasts for everything from agriculture to business, transport, and daily commute. It is crucial to accurately predict the weather to enable simple and seamless movement, as well as safe day to day operations. As the entire world is experiencing the effects of ongoing climate change.

The weather in a certain area can have a significant impact on the weather in other areas because weather systems can travel slowly across great distances in all directions. In this study, we present a technique to forecast weather conditions by aggregating historical weather data from a given place over one decade.

Using machine learning main goal is to create systems that can learn from their experiences and construct hypotheses as a result. Machine learning algorithms are trained on a training dataset to create a model. On the basis of the new input data, the model forecasts the weather. Using machine learning, it creates models from the data in the input dataset. It generates precise forecasts for clean datasets. following the processing of the dataset and the filling of any null values. The model is tested for accuracy using the fresh input data before making a weather prediction for the following day.

This research paper utilizes classification techniques to forecast the weather. Using this data, we can create straightforward machine-learning models that can predict the weather with high accuracy for the following day. These simple models can be executed on low-cost, low-resource computing platforms while still producing prompt and accurate forecasts that are applicable to daily life.

The major contributions of this paper include:

- The weather Prediction with machine learning.
- The testing of 3 machine learning algorithms.
- A thorough assessment of the proposed methodology,as well as a comparison of various time series models for forecasting future weather conditions. [1]. [2].

This paper is divided into 5 sections, related works, proposed methodology, results, and conclusion.

II. Related Work

The study of weather forecasting is not new; numerous data scientists have worked in this area and produced highly precise results. We will briefly discuss a few of the papers we read and evaluated to aid in our research, and we will cite all of them in the references section.

Mehmet Tektaş[3] analyses and contrasts the research on statistical and neuro-fuzzy network models for predicting the weather in Göztepe, Istanbul, Turkey. The average daily temperature (dry-wet), air pressure, and wind speed were calculated using 9 years worth of data. Using ANFIS and ARIMA time series forecasting models. 5% of the data in the ANFIS model were utilized for testing, while 9% were used for training. The ARIMA(2,1,1) approach evaluated performance criteria using the same data. ANFIS and ARIMA model performance comparisons based on MAE, RMSE, and R2 CTERA. Results revealed that ANFIS produced better results.

Salman and Kanigoro in their paper[4] recommended using the grid technique to forecast improved visibility for the varied values of the parameters p, d, and q. According to this experiment, ARIMA has the lowest MSE of 0.00029 and the lowest coefficient of variance of 0.00315. The MSE value in the ARIMA model rises with the number of prediction data.

in their paper, A Study of Time Series Models ARIMA and ETS [5] According to the authors, the difficulty of weather forecasting has increased as a result of shifting weather patterns. The research study uses the techniques ARIMA: autoagressive Integrated Moving Average and ETS: exponential smoothing to understand and analyzee the data of the specified parameters and observe the forecast for a short period of time.When comparing modeling techniques utilizing the ARIMA and ETS approaches, data from meteorological centers is used. Packages like ggplot2, forecast, time Date in R, and automatic prediction algorithms are also available inside these packages. According to accuracy, they usually try the simplest Methodology. On the basis of MAE, MASE, MAPE, and RMSE, their model will compare. The identification of the model will involve chromatic analysis of the ACF and PACF to make a variety of hypotheses. estimated by selection criteria AIC, AICc, and BIC.

Chen et al.[6] They analyzed temperature data from January 1951 to December 2017 to support their claim that employing SARIMA for seasonal time forecasting is an effective version of ARIMA. Chen, Niu, Liu, Jiang, and Ma. They claimed that the chosen SARIMA model can be utilized to estimate future values because its forecasting accuracy is adequate (training 1951–2014 and testing 2015–2017).

R. J. Boynton et al.[7] They recommended Reviewing the NARMAX approach's application to the Dst (disturbance storm time) index and the electron fluxes in geostationary Earth orbit (GEO), new information about the system's physics is revealed.

Imran Maqsood et al.[8] They demonstrate in this study how a group of artificial neural networks (ANNs) and learning paradigms can be used to forecast the weather in southern Saskatchewan, Canada. In comparison to existing methods like the linear combination, the suggested ensemble method for weather forecasting has advantages. It is also evident that the suggested neural networks outperform the traditional MLPN, ERNN, HFM, and RBFN models in both learning and generalization.

Ervin Zsótér et al.[9] The Extreme Forecast Index (EFI) was created to show instances where the EPS forecast distribution deviates significantly from the model climate (for additional information, see Lalaurette, 2002). It is regarded as helpful supporting data for other EPS products, including probability maps or EPSgrams.The major benefit of adopting the EFI is that it is an integral measure that is linked to the model climate and contains all the data related to a parameter's fluctuation in space and time. As a result, users are able to distinguish whether anomalies without having to set distinct space- and time-dependent thresholds.

Timann and Adrian et al.[10] Conclusion: Scoring criteria assign a numerical score based on the predictive distribution and the event or value that actually occurs to evaluate the accuracy of probabilistic forecasts. If the forecaster maximizes the anticipated score for an observation chosen from the distribution F if he or she gives the probabilistic forecast F, rather than G not equal F, then the scoring method is appropriate. If the maximum is unique, it is strictly appropriate. Proper grading methods for prediction tasks encourage the forecaster to make cautious judgments and to be sincere. Strictly correct scoring procedures in estimate problems offer desirable loss and utility functions that can be adjusted to the specific problem at hand.

[11] In order to meet the needs of the scientific and climate monitoring communities, the NCEP and NCAR are collaborating on a project (dubbed "reanalysis") to compile a 40-year record of worldwide assessments of atmospheric fields. In order to complete this task, data from the ground surface, ships, aeroplanes, satellites, rawinsondes, and other sources must be recovered, quality-controlled, and assimilated using a data assimilation system that has not changed throughout the reanalysis period (1957–1966). By doing this, perceived climate jumps caused by modifications to the data assimilation system are eliminated.

Leith has predicted in their paper et al.[12]that, It is discovered that the for an ideal situation with an observing resolution, perhaps possible in the 1980s with satellite-based sensors. The 6- to 10-day range is when the mean-square vector wind forecast skill improves the most with the Monte Carlo method. For a different situation that roughly corresponds to the current operational resolution, the quality of wind forecasting has significantly increased in the 2- to 5-day timeframe. The optimal filtering feature of the method, which dampens inaccurate small-scale structure in favor of the more predictable large scales, is responsible for a major portion of the improvement in mean-square skill.

[13] Forced dissipative hydrodynamic flow can be represented using deterministic ordinary nonlinear differential equations finite systems. Trajectories in phase space can be used to locate the solutions of these equations. Nonperiodic solutions are discovered to be typically unstable with respect to modest adjustments for those systems with bounded solutions, meaning that slightly different beginning states might evolve into significantly different ones. It is demonstrated that systems with bounded solutions also have bounded numerical solutions.

[14] Precipitation is directly impacted by global warming. Increased temperature results in more surface drying and evaporation, which increases drought intensity and duration. Most models predict precipitation that happens too early, too frequently, and with insufficient intensity, leading to an excessive amount of recycling and an inadequate amount of time for the moisture to stay in the atmosphere, which affects

runoff and soil moisture.

In their research [15] The findings indicate that there have been significant decadal or longer-term swings in the region's storminess conditions, with significant seasonal and regional variations. The North Sea area and other areas of the region exhibit the most pronounced changes between winter and summer. In the majority of this region, storminess appears to have decreased in the summer. The region's storminess pattern is characterized by rises in the north and decreases in the south during the transitional seasons, with increases in the north being greater in spring.

In their research [16] RMSE was 10.78, 11.20, 12.43, and 14.73 for the multivariate, univariate, and ARIMA and ARIMAX (Autoregressive Integrated Moving Average Model with Exogenous Input Variables) models, respectively. The performance of the LSTM model with exogenous meteorological data is the best of the four models, and meteorological factors can improve the LSTM model's forecast accuracy. Exogenous meteorological factors for the ARIMA model turned into an interference element rather than improving prediction accuracy.

In their research [17] For predicting short-term wind speeds, the Long Short-term Memory (LSTM) mode with deep learning capabilities in combination with fuzzy-rough set theory has been proposed. Reduced input and spatial properties are possible with fuzzy rough sets. The primary determinants of wind speed were identified as input for the LSTM neural network prediction model. Deep learning follows the big data trend. It is quite good at generalizing what it learns from large amounts of data. The experimental findings demonstrate that the Fuzzy Rough Set Long Short-Term Memory (FRS-LSTM) model outperforms the conventional neural network in terms of prediction accuracy.

In their research [18] Probability integral transform (PIT) histograms can be used to verify probabilistic calibration in real-world settings. The use of appropriate scoring criteria, such as the logarithmic score and the continuously ranked probability score, allows for the simultaneous evaluation of calibration and sharpness.

In their article, [19] Historical sun irradiance data, correlations between different meteorological variables (such as wind speed, humidity, and cloudiness), and interactions between the weather contexts of physically adjacent places are all necessary for accurate forecasting.

In their article, [17] The root mean square error, mean absolute error, mean absolute percentage error, and symmetric mean absolute percentage error of the proposed model, according to numerical results, are respectively 0.2047, 0.1435, 3.77%, and 3.74%, outperforming benchmark predictions made using well-liked parameter optimization techniques, data processing methods, and hybrid neural network forecasting models.

In their article, [20] In this research, three different kinds of goodness are identified. 1) the consistency, or type 1 goodness, between forecasters' assessments and their predictions; 2) the quality, or type 2 goodness, or matching observations; and 3) the additional financial and/or non-financial benefits that decision-makers derive from using the forecasts, or type 3 goodness, or value. Each kind of goodness is defined and briefly explained. The measuring of consistency, quality, and value is also a topic of discussion.

In their article, [21] A functional regression model variant that could identify weather trends were utilized in addition to a linear regression model. Professional weather forecasting services beat both of our models, but the difference between our models and the professionals' performance quickly shrank for forecasts made for later days. It is possible that our models could even outperform professional ones for longer time scales. The functional regression model performed better than the linear regression model, indicating that two days were too short for the latter to capture significant weather trends. Perhaps the functional regression model would perform better if our forecasts were based on weather data for four or five days.

III. Proposed Methodology

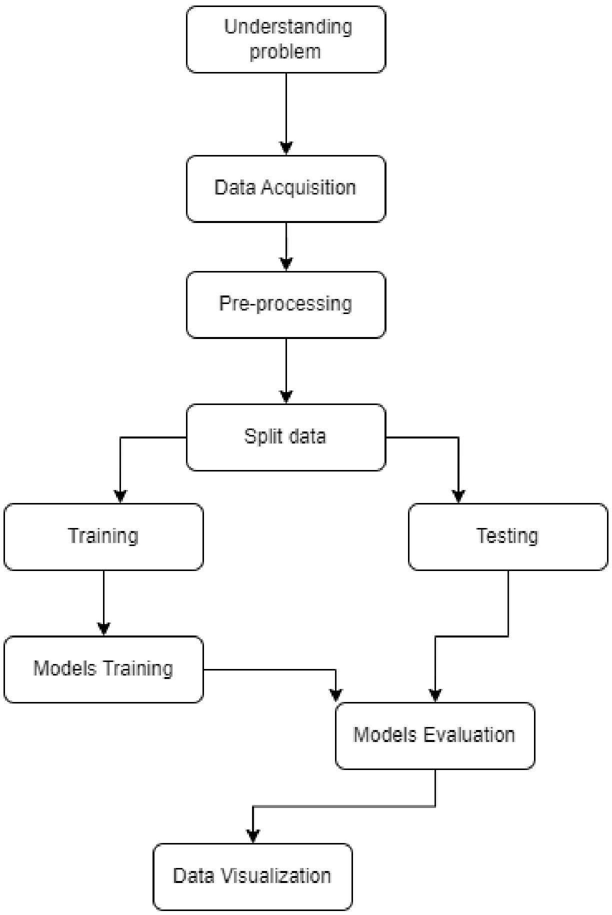

Fig. 1. Weather Forecasting process

A. *Dataset Description*

Our dataset consists of 9 features, and it holds 96,545 records. The dataset was split into two partitions: 80% for training, and 20% for testing. A detailed description of the features can be found below.

The Formatted Date appears in the first column to indicate which day and hour, specifically, the upcoming features measurements belong to. Precip Type, which stands for precipitation type and denotes whether the day features rain, snow, hail, or sleet, is the abbreviation. The amount of heat in the air is gauged by temperature. When relative humidity and air temperature are combined, the outcome is what is known as apparent temperature, which is how the temperature feels when it interacts with the human body. The 'temperature' and 'apparent temperature' features both use degrees Celsius as their units of measurement. The amount of water vapour carried in the air is estimated by the humidity, which provides a normalised number between 0 and 1. Wind Speed displays the velocity of the wind in kilometres per hour. The direction of the wind according to a 360-degree measurement perspective is known as wind bearing. The horizontal distance at which a person should be able to see and recognise is stated in kilometres and serves as a measure of the horizontal opacity of the atmosphere at the site of observation. Millibars are used to measure pressure.

Sorting these traits from most crucial to least crucial. We decide to focus on the top four most important characteristics moving forward. Temperature, Humidity, Wind Speed, and Visibility are the features that have been picked. To fit the machine learning models, we will first divide the data into training and testing sets. This will enable us to forecast the four features for a desired date and time in the future.

TABLE I
FEATURES OF DATASET

Features	Type	Values
Formatted Date	Numerical	From 1Jan06 to 1Jan17
PrecipType	Classification	Rain or Snow
Temperature	Numerical	From -22 to 40
Apparent Temperature	Numerical	From -28 to 40
Humidity	Numerical	From 0 to 1
Wind Speed	Numerical	From 0 to 64
Wind Speed	Numerical	From 0 to 360
Visibility	Numerical	From 0 to 16
Daily Summary (Target)	Classification	Result of Temp

B. *Dataset Visualization*

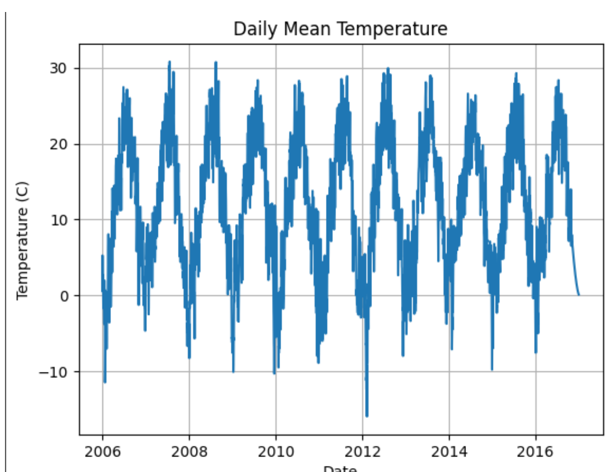

Fig. 2. Visualization of Temperature

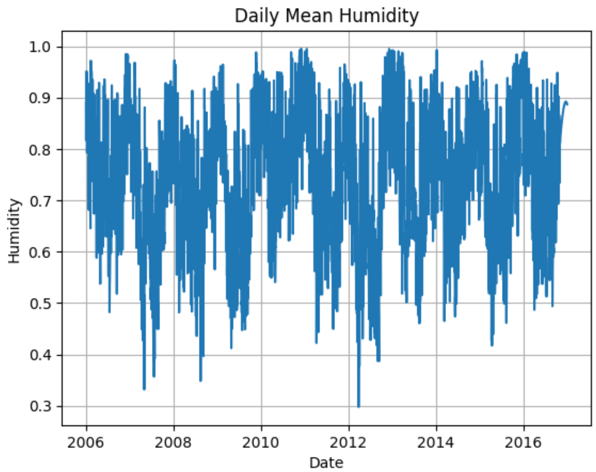

Fig. 3. Visualization of Humidity

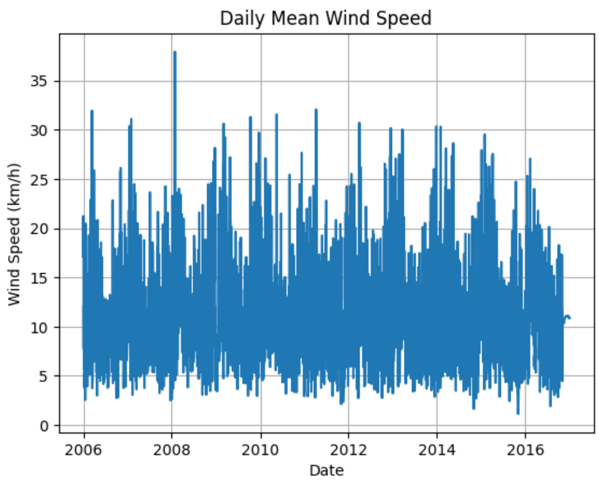

Fig. 4. Visualization of Wind Speed

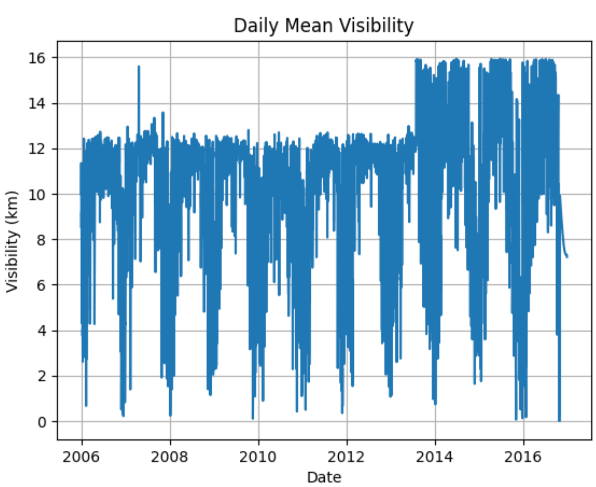

Fig. 5. Visualization of Visibility

C. *Used Algorithms*

Four distinct machine learning algorithms, ARIMA (Autoagressive intergrated moving average model), SARIMA (Seasonal autoagressive inegrated moving average model), and SARIMAX (seasonal autoagressive integareted moving average model using Exogenous factors) .were applied to the aforementioned dataset with the selected attributes. These data included Mean Absolute Error, Mean Square Error, Root Mean Squared Error, Accuracy, Recall, Precision, and Specificity for each algorithm. The outcomes were then compared and recorded. You can find the results, charts, and a discussion of the results later in the paper.

1) **Autoregressive Integrated Moving Average Model:**

A statistical analysis model called an autoregressive integrated moving average, or ARIMA uses time series data to either better comprehend the data set or forecast future trends. If a statistical model forecasts future values using data from the past, it is said to be autoregressive. [22].

$$\varphi_p(G)\varphi_p\left(G^*\right)(1-G)^d\left(1-G^*\right)^D X_t=\gamma_q(G) w_Q\left(G^*\right) e_t \tag{1}$$

2) Seasonal Autoregressive Integrated Moving Average Model:

Seasonal SARIMA ARIMA, or more formally, ARIMA with a seasonal element. SARIMA is a popular time series analysis method for forecasting future values using previous data with a seasonal component. As an illustration, consider the seasonal sales of electrical equipment and the long-term weather forecast.[23].

$$\varphi_p(G)\varphi_p\left(G^s\right)(1-G)^d\left(1-G^s\right)^D X_t=\alpha_k y_{k,t}+\gamma_q(G) w_Q\left(G^s\right) e_t \tag{2}$$

3) Seasonal Autoregressive Integrated Moving Average Model with exogenous factors:

The SARIMAX model is an upgraded version of the SARIMA model that incorporates exogenous elements (X) as external feature parameters to improve model performance, lower prediction errors, address autocorrelation concerns, and produce better prediction outcomes [45]. The exogenous factors are optional parameters in the SARIMAX model, which includes seasonal effects and exogenous factors that can be employed as SARIMAX (p, d, q) (P, D, Q). Exogenous factors can be external parallel time series data that have the same correlation as the source data and need to be anticipated, such as wind speed or temperature measurements. The prediction mode is supported by the external components.l[24].

$$\begin{aligned} e_t &= \frac{\gamma(G)}{\varphi(G)} a_t \\ y_t &= \alpha+\sum_{i=1}^{m} \frac{\gamma_i(G)}{\varphi_i(G)} G l_i X_t+e_t \end{aligned} \tag{3}$$

IV. Results and Analysis

By calculating the errors of each Model used with all of the features, we have calculated the relation between our three models and temperature, humidity, wind speed, and visibility. Using MAE, MSE, and RMSE.

$$\text{Mean Absolute Error} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_1|}{n} \quad (4)$$

$$\text{Mean Square Error} = \frac{1}{N} \sum_{i=1}^{N} (f_i - y_i)^2 \quad (5)$$

$$\text{Root Mean Square Error} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (y_j - \hat{y}_j)^2} \quad (6)$$

The results collected from ARIMA, SARIMA, SARIMAX are shown beneath. The following results are from our dataset.

The following results are from the first dataset.

TABLE II
STATISTICS OF THE PREDICTED FEATURES USING MEAN ABSOLUTE ERROR

Model	ARIMA	SARIMA	SARIMAX
Temperature	1.5375	1.4933	1.3531
Humidity	0.06257	0.06251	0.0620
Wind Speed	2.8051	3.2808	3.1892
Visibility	2.9731	1.6348	1.5710

By using Mean Absolute Error here, it shows that for the Temperature the best model is SARIMAX, for Humidity the best model is SARIMAX, for Wind Speed the best model is ARIMA, and for Visibility the best model is SARIMAX.

TABLE III
STATISTICS OF THE PREDICTED FEATURES USING MEAN SQUARE ERROR

Model	ARIMA	SARIMA	SARIMAX
Temperature	4.7093	4.2311	3.4599
Humidity	0.0065	0.00924	0.00920
Wind Speed	14.6837	19.7344	17.4293
Visibility	13.4545	5.8349	5.3565

By using Mean Square Error here, it shows that for the Temperature the best model is SARIMAX, for Humidity the best model is ARIMA, for Wind Speed the best model is ARIMA, and for Visibility the best model is SARIMAX.

TABLE IV
STATISTICS OF THE PREDICTED FEATURES USING ROOT MEAN SQUARE ERROR

Model	ARIMA	SARIMA	SARIMAX
Temperature	2.17010	2.0569	1.8600
Humidity	0.08095	0.09615	0.0959
Wind Speed	3.8319	4.4423	4.1748
Visibility	3.6680	2.4155	2.3144

By using Root Mean Square Error here, it shows that for the Temperature the best model is SARIMAX, for Humidity the best model is ARIMA, for Wind Speed the best model is ARIMA, and for Visibility the best model is SARIMAX.

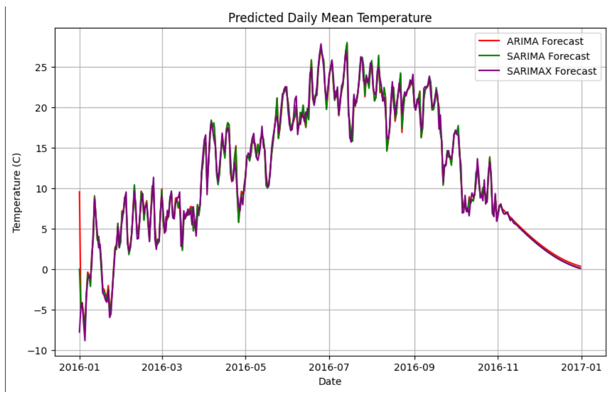

Fig. 6. Visualization of Temperature Predicted Data

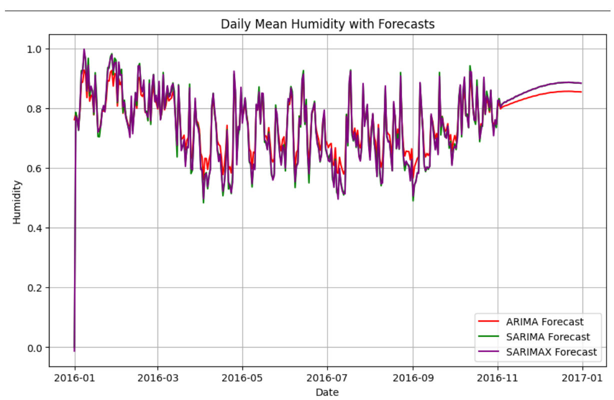

Fig. 7. Visualization of Humidity Predicted Data

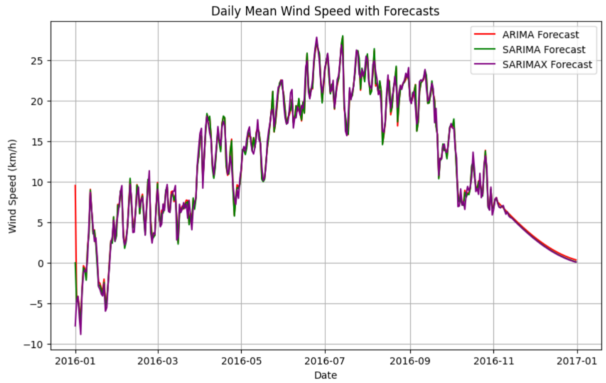

Fig. 8. Visualization of Wind Speed Predicted Data

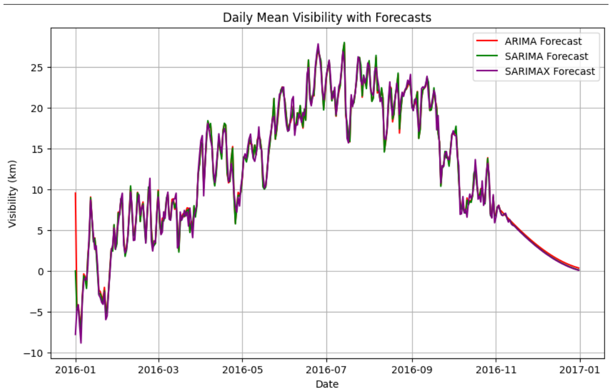

Fig. 9. Visualization of Visibility Predicted Data

V. Conclusion

It is crucial to forecast the weather, and machine learning has proven to be a wonderful method for doing so. The prediction accuracy could be increased even more with additional study and advice from weather experts. The weather models ARIMA, SARIMA, and SARIMAX have been shown to be extremely accurate with only small error values. The ability of machine learning to predict outcomes of events like sporting contests and stock prices across a wide range of industries—not just meteorology—makes it a very useful tool for humanity. And this technology will just keep improving and producing greater outcomes.

References

[1] S. K. Jayasingh, "Smart weather prediction using machine learning," *Research Gate*, 2022.

[2] A. H. M. Jakaria, "Smart weather forecasting using machine learning:a case study in tennessee," *Research Gate*, 2022.

[3] M. Tektaş, "Weather forecasting using anfis and arima models. a case study for istanbul," vol. 51, 2010.

[4] B. K. Afan Galih Salmana, "Visibility forecasting using autoregressive integrated moving average (arima) models," *ScienceDirect*, 2020.

[5] B. M. Garima Jain, "A study of time series models arima and ets," *SSRN*, 2017.

[6] D. L.-W. J. Peng Chen1, Aichen Niu1 and B. Ma1, "Time series forecasting of temperatures using sarima: An example from nanjing," vol. 394, 2018.

[7] S. A. B. R. J. Boynton, M. A. Balikhin and O. A. Amariutei, "Application of nonlinear autoregressive moving average exogenous input models to geospace: advances in understanding and space weather forecasts," *EGU*, 2013.

[8] M. R. K. A. A. Imran Maqsood, "An ensemble of neural networks for weather forecasting," 2004.

[9] E. Zsótér, "Recent developments in extreme weather forecasting," 2006.

[10] T. GNEITING and A. E. RAFTERY, "Strictly proper scoring rules, prediction, and estimation," *American Statistical Association*, 2007.

[11] E.Kalnay, "he ncep/ncar 40-year reanalysis project. bulletin of the american meteorological society," *AMS*, vol. 77, pp. 437–472, 1996.

[12] C. E. Leith, "Theoretical skill of monte carlo forecasts," *AMS*, vol. 102, no. 6, 1974.

[13] E. N. Lorenz, "Deterministic nonperiodic flow," 1963.

[14] K. E. Trenberth*, "Changes in precipitation with climate change," vol. 47, 2011.

[15] X. L. W. F. W. Z. V. R. S. Y. Feng, "Trends and variability of storminess in the northeast atlantic region, 1874–2007," 2008.

[16] . Y. M. . S. W. . S. L. . R. N. . Y. W. . Q. G. . . Rui Zhang, 1 Zhen Guo and . Yonghong Li4, "comparison of arima and lstm in forecasting the incidence of hfmd combined and uncombined with exogenous meteorological variables in ningbo, china," *pybmed central*, 2021.

[17] W. W. L. P. Jingchun Chu, Ling Yuan, "Short-term wind speed prediction based on deep learning," 2019.

[18] T. Gneiting1 and M. Katzfuss2, "Probabilistic forecasting," *annual review*, 2014.

[19] M.-W. C. . yeon-Ju Jeon, 1 and . O-Joun Lee2, "Day-ahead hourly solar irradiance forecasting based on multi-attributed spatio-temporal graph convolutional network," 2022.

[20] A. H. Murphy, "What is a good forecast? an essay on the nature of goodness in weather forecasting," *AMS*, vol. 8, 1993.

[21] C. V. Mark Holmstrom, Dylan Liu, "Machine learning applied to weather forecasting," 2016.

[22] A. HAYES, "Autoregressive integrated moving average (arima) prediction model," *Investopedia*, 2022.

[23] P. AGGARWAL, "Sarima using python – forecast seasonal data," *Wisedomgeek*, 2020.

[24] by Fahad Radhi Alharbi *Denes Csala, "A seasonal autoregressive integrated moving average with exogenous factors (sarimax) forecasting model-based time series approach," *MDPI*, 2022.