Weather Forecasting using Machine Learning

SK SAMIUL KADIR
School of Data and Sciences
Brac University
Dhaka, Bangladesh
sk.samiul.kadir@g.bracu.ac.bd

***Abstract*—The ability to predict the state of the atmosphere at a specific area and time is made possible through the use of scientific methodologies and technology in weather forecasting. It's one of the world's most challenging problems. To determine which machine learning models are most effective at forecasting the weather, this study uses a variety of models, including ARIMA, LSTM, prophet,RandomForestRegressor and Var. The output of multiple models is also examined and contrasted using various errors between the anticipated and actual numbers as well as graphing.. This work makes use of Pandas, NumPy, Keras, Git, and Matplotlib.**

I. Introduction

The use of scientific methods and technology in weather forecasting enables the prediction of the atmosphere's state at a certain location and moment. It's one of the most difficult issues all over the world. The project aims to solve this problem by using Machine Learning Models. In the past, weather forecasting was done manually using variations in barometric pressure, the current state of the atmosphere, and the state of the sky or cloud cover. Today, weather forecasting is done using computer-based models that take numerous atmospheric factors into account. The researcher had spent a lot of effort trying to create a linear link between the attributes of the raw meteorological data and the matching target attribute.

II. Related Words

Weather Forecast Prediction: An Integrated Approach for Analyzing and Measuring Weather Data, this Paper aims to use different techniques to estimate the weather. [1] According to them, Weather forecasting is an issue worldwide and needs to be done accurately. That's why they are trying o tackle his problem. They used Naive Bayes and Chi-Square Algorithms for classification purposes. Their system is a web Application with an effective graphical user interface. It takes a specific amount of parameters and predicts the weather. Furthermore, they achieved an accuracy of up to 97 in prediction. I think, more attributes of weather conditions can be handled to predict and use another classification algorithm for better accuracy.

The ARIMA (Autoregressive Integrated Moving Average) model is the main topic of the study titled "Forecasting of Demand Using the ARIMA Model." [2]

The authors want to address the difficulty of correctly estimating future demand in a variety of industries, including supply chain management, manufacturing, and retail.

To handle non-stationary time series data, they use the ARIMA model, which combines autoregressive and moving average components with differencing. Additionally, ARIMA is a very well-liked model for forecasting and time series data. However, it could be difficult to locate precise parameters (p, d, q). The ARIMA model makes predictions based on historical data or observations. ACF and PACF are necessary to find the parameter's proper values. The strategy entails fitting the ARIMA model to historical demand data and utilizing it to predict future demand based on the discovered patterns and trends. The authors offer insights into demand forecasting by utilizing the ARIMA model, supporting firms in streamlining inventory management, production planning, and resource allocation.

Using a Convolutional LSTM (ConvLSTM) network, the paper "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting" suggests a novel approach for precipitation nowcasting. [3]

The authors use spatiotemporal correlations in weather radar data to address the issue of precisely forecasting short-term precipitation.

In order to capture spatial interdependence and temporal dynamics in the radar data, they develop the ConvLSTM architecture, which combines the strengths of convolutional neural networks (CNNs) with LSTM (Long Short-Term Memory) networks.

The method involves generating forecasts for upcoming precipitation at precise temporal intervals by training the ConvLSTM network on past radar recordings.

ConvLSTM network outperforms conventional algorithms and produces encouraging results in precipitation nowcasting, proving the value of deep learning techniques for spatiotemporal forecasting applications.

A Data Mining Paradigm to Forecast Weather-Sensitive Short-Term Energy Consumption First, if we want to know about Data Mining, it's the process where we discover patterns and trends of a large dataset. The authors of this paper try to tackle the problem of accurately predicting energy consumption or to be precise short-time energy consumption and the data is based on weather. To tackle this they used the

data-mining paradigm which is a particular way of thinking about a particular field. This data mining combines different models which give a prediction, and feature selection to forecast energy consumption. [4] Overall it's a better approach for accurately predicting short-term energy consumption than other studies which had been made until now.

A method for weather prediction using machine learning techniques is presented in the work titled "Smart Weather Prediction Using Machine Learning".

By utilizing machine learning methods, the authors hope to address the problem of properly anticipating meteorological conditions. [5]

To forecast meteorological variables like temperature, humidity, and precipitation, they investigate several machine learning methods, such as regression, random forests, and neural networks.

In order to anticipate the weather in the future, this method includes training these models on historical weather data.

The authors want to use machine learning to increase the dependability and accuracy of weather forecasting, which will help with better decision-making for a variety of applications like disaster management, transportation, and agriculture.

MACHINE LEARNING TECHNIQUES FOR WEATHER FORECASTING, this paper mainly reviews already made different techniques by different Machine learning processes. The main problem they try to solve is the challenge of forecasting weather or predicting weather beforehand. [6]

The author of this paper discusses different Machine Learning algorithms to predict the weather. Like, Decision Tree was used in the prediction of cyclones from different features available. Furthermore, the author mentioned how the availability of data can be a limitation of Machine Learning. According to me, this paper identifies the pros and cons of ML and gave suggestions about the idea of continuing to do research in this area.

The article "Smart Weather Forecasting Using Machine Learning: A Case Study in Tennessee" examines how machine learning techniques were used to forecast the weather in Tennessee. [7]

The authors use machine learning techniques that are region-specific to address the problem of increasing weather prediction accuracy.

To forecast weather variables like temperature and precipitation, they investigate different machine learning models, such as random forests and support vector regression.

Using historical weather data from Tennessee, these models are trained and then fine-tuned. Their performance is then compared to that of more established forecasting techniques.

The authors want to improve weather forecasting capabilities by utilizing machine learning. This would enable more accurate and trustworthy predictions for the area, which can have substantial ramifications for industries like agriculture and energy management.

The article "Smart Weather Forecasting Using Machine Learning: A Case Study in Tennessee" examines how machine learning techniques were used to forecast the weather in Tennessee. The authors use machine learning techniques that are region-specific to address the problem of increasing weather prediction accuracy. [8] To forecast weather variables like temperature and precipitation, they investigate different machine learning models, such as random forests and support vector regression. Using historical weather data from Tennessee, these models are trained and then fine-tuned. Their performance is then compared to that of more established forecasting techniques. The authors want to improve weather forecasting capabilities by utilizing machine learning. This would enable more accurate and trustworthy predictions for the area, which can have substantial ramifications for industries like agriculture and energy management.

The author of the study "Machine Learning Applied to Weather Forecasting" points out the drawbacks of conventional forecasting methods, which frequently have long-term mistakes. The goal of the study is to increase the precision of long-range weather forecasting in order to overcome this. The two machine learning (ML) model types that the author primarily uses are linear regression and functional regression models. These models contain extra characteristics like temperature, pressure, and wind speed to increase accuracy and are trained using historical data. [9] The outcomes show that the ML models suggested in the research perform more accurately than conventional forecasting models. However, it is important to remember that the study's geographic focus is restricted. Overall, the research emphasizes the potential of ML approaches in weather forecasting, especially for longer-term predictions, and emphasizes the necessity of including extra characteristics to increase accuracy. Stronger and more dependable weather prediction systems might be developed as a result of additional ML study and investigation in various geographical areas.

The subject of increasing forecasting accuracy by applying business cycle limits on Vector Autoregressive (VAR) models is addressed in the study "Forecasting with Vector Autoregressive (VAR) Models Subject to Business Cycle Restrictions" While taking into account the cyclical structure of the economy, the authors seek to represent the dynamic interactions between economic variables. [**?**] The strategy entails using constraints on parameter estimate to introduce prior knowledge of business cycles into VAR models. These limitations enable the models to more accurately depict cyclical patterns and interactions between variables, producing more precise projections. The paper adds to the body of knowledge on VAR modeling by highlighting the value of business cycle constraints for forecasting precision. To prove the viability of the suggested strategy,

it offers comparative analysis and empirical proof. The results indicate that adding business cycle limits enhances the accuracy of VAR models when used for economic forecasting.

The study "Time Series Forecasting of Temperatures Using SARIMA: An Example from Nanjing" focuses on temperature forecasting in the setting of Nanjing using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model.

By using SARIMA modeling techniques on historical temperature data, the authors tackle the problem of precisely forecasting temperature changes. The goal is to create a forecasting model that is accurate and captures seasonal trends and patterns in temperature data. [10] The strategy entails choosing the right model parameters for the SARIMA model using assessment and model selection methods. To create a reliable forecasting model, the authors take into account the seasonal and trend components of the temperature time series. [11] By proving the use of SARIMA models for temperature forecasting, specifically in the setting of Nanjing, the research adds to the body of literature. The findings shed light on how SARIMA models can be used for climate studies and assist in guiding decision-making in fields like agriculture, energy, and urban planning that are susceptible to temperature changes.

The paper "An Introductory Study on Time Series Modeling and Forecasting" offers a thorough introduction of time series modeling and forecasting methods. [12] The authors want to familiarize readers with the core ideas, procedures, and strategies employed in time series analysis. They cover several models, highlighting their advantages and disadvantages, such as the autoregressive (AR), moving average (MA), and autoregressive integrated moving average (ARIMA) models. The issue of predicting future values in time series data is also covered in the study, with a focus on model selection, parameter estimates, and assessment metrics. The survey of the literature explores various applications and approaches put forth by scholars while highlighting significant contributions to time series analysis and forecasting. Overall, the paper acts as a primer on time series modeling and forecasting, giving readers a basic overview of the subject and providing references to further investigate cutting-edge methods in the area.

A thorough analysis of the state of artificial neural networks (ANNs) in the field of forecasting is given in the publication "Forecasting with Artificial Neural Networks: The State of the Art". [13] The objective of the authors is to review and compile the developments in ANNs for forecasting applications, including time series forecasting. They cover numerous ANN architectures, including feedforward neural networks, recurrent neural networks, and convolutional neural networks, highlighting their advantages and disadvantages in various forecasting scenarios. In order to improve forecasting accuracy and identify intricate patterns and relationships in data, the paper makes use of ANNs' computational power and learning skills. The literature review discusses the benefits, drawbacks, and probable future directions of ANNs in forecasting while offering insights into recent advancements and applications. Overall, the paper offers state-of-the-art methods and breakthroughs in the field, serving as a great resource for scholars and practitioners interested in using ANNs for predicting jobs

The paper "Time Series Forecasting Using Hybrid ARIMA and ANN Models Based on DWT Decomposition" puts forth a hybrid method for time series forecasting that combines Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN) models based on Discrete Wavelet Transform (DWT) decomposition. [14] By combining the benefits of both the ARIMA and ANN models, the authors hope to increase the precision of time series forecasting. To capture the underlying patterns at various scales, they deconstruct the time series using DWT, and they then use ARIMA and ANN models to forecast each component. In particular, when non-linear and non-stationary trends are present, the study addresses the challenge of effectively collecting and predicting complex patterns and changes in time series data. The survey of the literature gives a general overview of time series forecasting techniques and procedures while noting the shortcomings of individual models and the demand for hybrid strategies that integrate several methodology.

Overall, the research introduces a novel hybrid strategy to enhance the precision and resilience of time series forecasting by combining the strengths of ARIMA, ANN, and DWT decomposition.

A thorough review of time series analysis methods for predicting and control is given in the book "Time Series Analysis: Forecasting and Control, 5th Edition." [15]

For the analysis and modeling of time series data, the authors discuss a variety of strategies and approaches, including traditional time series models, state-space models, and cutting-edge forecasting methods.

The goal of the book is to help practitioners and academics make well-informed judgments based on precise forecasts by addressing the issue of comprehending and forecasting the behavior of time-dependent data.

The authors walk readers through the time series modeling, forecasting, and control processes using a blend of theoretical concepts and real-world examples, emphasizing the significance of choosing the right models and applying the right methods to various types of time series data.

III. Research Methodology

Our research aims to compare various machine learning techniques for weather prediction. These techniques utilize features such as mean temperature, wind speed, humidity, and pressure to forecast weather conditions. To train our machine learning models, we collected weather data spanning

from 1901 to 2017. In our study, we employ a range of models including ARIMA, DecisionTreeRegressor, LSTM, KNN, Prophet, and VAR. Through evaluating and training these models, we seek to assess their performance and identify the most effective approach for weather forecasting.

A. Data-Preprocessing

Here we used two datasets which are from Kaggle. Data cleaning entails dealing with missing data, eliminating duplicates, and addressing outliers. Depending on the situation, missing data are eliminated or imputed. To maintain data integrity, duplicates are found and removed.

Data transformation entails transforming data into a format that is appropriate for analysis. Numerical features may need to be scaled using MaxScaler or normalized to a common range, categorical variables need to be encoded into numerical representations. Data were made stationary for specific MODELS. Converted date column to datetime format.

B. Feature Scalling

The most relevant traits must be chosen, or new features must be extracted from the current ones, in this process. It contributes to dimensionality reduction, model performance improvement, and interpretability improvement. Methods like correlation analysis were used.

C. Using Different Models of ML

1) ARIMA: The ARIMA model is a popular and effective approach for time series forecasting across various domains. Its widespread use and success in numerous applications make it a reliable choice for forecasting tasks. By applying the ARIMA model to historical demand data, it becomes a valuable tool for predicting future temperatures or other variables that vary over time. This modeling technique allows for capturing patterns and trends from the past to make accurate predictions about future values. Hence, leveraging the ARIMA model on historical data provides a robust framework for forecasting time-dependent variables, enabling informed decision-making in various industries and domains.

To develop predictions based on past data or observations, the ARIMA (Autoregressive Integrated Moving Average) model combines autoregressive (AR), integrated (I), and moving average (MA) components. The I component deals with the integration of the data to achieve stationarity, while the MA component captures the dependence between the variable and its residual errors. The AR component represents the linear relationship between the variable and its lagged values.

We used the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots to determine the appropriate parameter values (p, d, and q) for the ARIMA model. By looking at the major lag values, the ACF plot aids in determining the MA order (q), while the PACF plot aids in determining the AR order (p), by looking at the significant partial autocorrelation values. To assess the accuracy of the predictions, I employed the meanSquaredError metric, which provided an RMSE (Root Mean Squared Error) value of 3.90. The RMSE is a measure of the average difference between the predicted values and the actual values. A lower RMSE indicates a better fit of the model to the data, suggesting that the predictions are relatively close to the true values. In this case, the obtained RMSE value of 3.90 suggests that the model's predictions exhibit a moderate level of accuracy.

The equation we used:

$$(1 - \Phi_1 L - \Phi_2 L^2 - \ldots - \Phi_p L^p)(1 - L)^d Y_t = \mu + (1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q) \quad (1)$$

Where:
y_t represents the time series data at time t.
L is the lag operator, which shifts the time series by one time step.
μ is the mean of the time series.
ϕ_1 are the autoregressive coefficients.
θ_1 are the moving average coefficients.
ϵ_t represents the white noise or error term at time t.

2) LSTM: LSTM, which stands for Long Short-Term Memory, is a type of recurrent neural network (RNN) model that is particularly effective in capturing patterns in sequential data. In our study, we chose LSTM because it can leverage the previous values in a time series to make accurate predictions. To feed the data into the LSTM model, we created batches of 12 input values using the TimeSeriesGenerator. During the training process, we monitored the loss function for each epoch and observed that the loss started to stabilize around 8 epochs. We visualized the predicted values alongside the actual data by plotting the mean temperature(Fig-1). To assess the accuracy of our predictions, we used the meanSuaredError metric, which yielded a value of 6.99

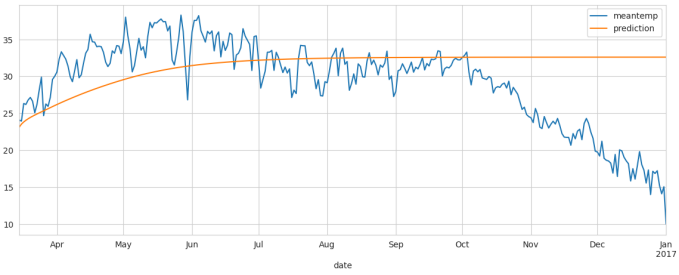

Fig. 1. Actual vs Prediction

3) Prophet: Prophet, a useful forecasting model we used in our study, was a key component of the analysis of our second dataset together with other models. It is unique in that it can handle a variety of seasonal patterns, including weekly and monthly fluctuations, automatically recognizing and including them into the projections. Additionally, Prophet successfully catches both transient variations and persistent trends found in the data. After using historical data to train the model, we used the built-in visualization features to evaluate how well

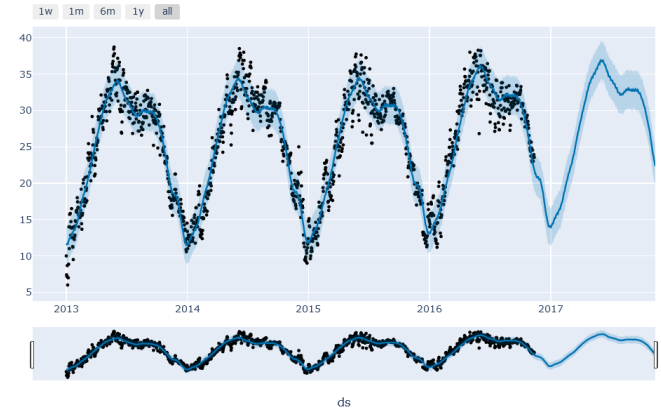

Fig. 2. Actualvs Prediction

the predictions matched the actual data.(Fig-2) The accuracy of the model was also determined by calculating the root mean squared error (RMSE), which yielded a result of around 9.275.

4) VAR: Multiple time series variables with interdependencies can be analyzed and predicted using the VAR (Vector Autoregressive) model. VAR considers the lagged values of many variables as opposed to the AR model, which only takes into consideration the prior values of one variable.

From the equation it'd be much clear:

$$Y_{1,t} = c_1 + A_{11}Y_{1,t-1} + A_{12}Y_{2,t-1} + \varepsilon_{1,t} \tag{2}$$

$$Y_{2,t} = c_2 + A_{21}Y_{1,t-1} + A_{22}Y_{2,t-1} + \varepsilon_{2,t} \tag{3}$$

The Akaike Information Criterion (AIC) is frequently used to choose the right lag order for the VAR model. We can choose the lag order that results in the lowest AIC value, indicating the best model fit, by fitting the VAR model with several lag orders. We have determined that lag order 21 in our code has the lowest AIC value. In order to guarantee stationarity, we then fit the VAR model with an order of (21, 0). We can produce forecasts for the desired number of future time points after the model has been fitted. Furthermore for accuracy check, we got an RMSE value of 5.4 and an MAE value of 4.31.

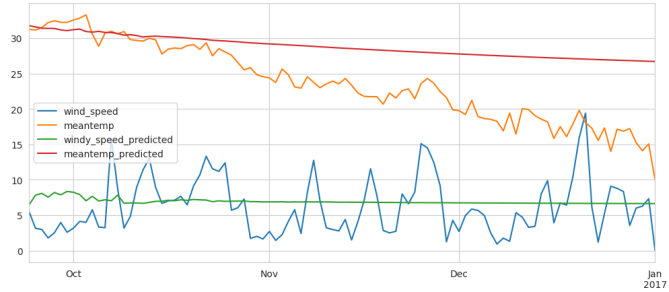

Fig. 3. Actualvs Prediction

5) RandomForestRegressor: Random Forest is a machine-learning model that utilizes multiple decision trees for prediction. It effectively addresses the issue of overfitting, which is commonly associated with decision trees. Overfitting happens when a model excessively memorizes the training data instead of capturing general patterns. Random Forest tackles this problem by training numerous trees on different subsets of the data and then combining their predictions.

Moreover, Random Forest helps mitigate the problem of bias, which arises when data is not evenly distributed among different classes or categories. By aggregating the predictions of multiple decision trees, Random Forest reduces the impact of bias and provides more robust and accurate predictions.

In my particular case, I employed the Random Forest Regressor due to its ability to handle overfitting and bias. I prepared the data by performing necessary preprocessing steps to ensure its suitability for the model. This involved creating appropriate input and output variables in a supervised learning setup.

IV. Results

Residual error, also known as the prediction error or the difference between the actual and predicted values, represents the unexplained variability in the data by the model. It is a fundamental concept used in various evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

MAE (Mean Absolute Error) calculates the average of the absolute values of the residuals. It provides a measure of the average magnitude of the errors without considering their direction. MAE is less biased for higher values because it treats all errors equally, regardless of their magnitude. It provides a balanced prediction error metric and is commonly used when outliers or extreme values are present in the data.

MSE (Mean Squared Error) calculates the average of the squared residuals. It amplifies larger errors due to the squaring operation, making it more sensitive to outliers or extreme values. MSE is more biased towards higher values because it penalizes larger errors more heavily. This can make it suitable for scenarios where larger errors need to be emphasized or when optimizing the model using gradient-based optimization methods.

RMSE (Root Mean Squared Error) is the square root of MSE and represents the average magnitude of the residuals. It is commonly used to have the error metric on the same scale as the original data. RMSE shares similar properties with MSE in terms of bias towards higher values.

The choice of the appropriate error metric depends on the specific context, objectives, and preferences of the problem. MAE is often preferred when a balanced prediction error is desired, while MSE and RMSE are commonly used when larger errors need to be penalized or when the error needs to be measured on the same scale as the data. Based on our plots, it is evident that the RandomForestRegressor produced lower RMSE and MAE values compared to other models.

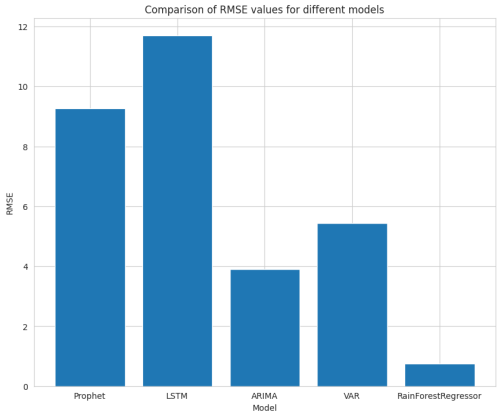

Fig. 4. Comparison of different RMSE

Additionally, the LSTM model performed the worst in terms of RMSE and MAE.

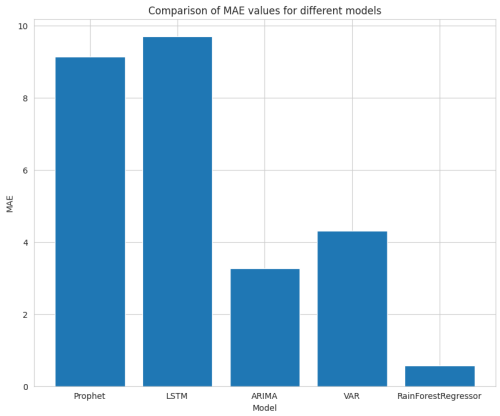

Fig. 5. Comparison of differnet MAE values

V. Conclusion

Weather forecasting, a complex task that relies on scientific methods and technology, aims to predict atmospheric conditions at specific locations and times. It poses significant challenges worldwide. To identify the most effective machine learning models for weather forecasting, this study employs various approaches such as ARIMA, LSTM, prophet, RandomForestRegressor, and Var. In addition to implementing one model on an existing dataset, an additional dataset was introduced, and multiple models were utilized to compare their performance. Ultimately, ARIMA and RandomForestRegressor yielded promising results.

References

[1] M. Biswas, T. Dhoom, and S. Barua, "Weather forecast prediction: An integrated approach for analyzing and measuring weather data," *International Journal of Computer Applications*, vol. 182, pp. 20–24, 12 2018.

[2] J. Fattah, L. Ezzine, Z. Aman, H. Moussami, and A. Lachhab, "Forecasting of demand using arima model," p. 184797901880867, 10 2018.

[3] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. K. Wong, and W.-c. WOO, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," 06 2015.

[4] M. Torabi and S. Hashemi, "A data mining paradigm to forecast weather sensitive short-term energy consumption," 05 2012, pp. 579–584.

[5] S. Jayasingh, J. Mantri, and S. Pradhan, *Smart Weather Prediction Using Machine Learning*, 05 2022, pp. 571–583.

[6] W. S. Sanders, "Machine learning techniques for weather forecasting," B.S. thesis, Florida State University, 2005.

[7] A. H. M. Jakaria *et al.*, "Smart weather forecasting using machine learning: A case study in tennessee," 08 2020, available at SSRN: https://ssrn.com/abstract=xxxxxxx.

[8] A. Patel, P. K. Singh, and S. Tandon, "Weather prediction using machine learning," February 2021, available at SSRN: https://ssrn.com/abstract=3836085 or http://dx.doi.org/10.2139/ssrn.3836085.

[9] M. Holmstrom, D. Liu, and C. Vo, "Machine learning applied to weather forecasting," *Stanford University*, December 2016, dated: December 15, 2016.

[10] P. Chen, A. Niu, D. Liu, W. Jiang, and B. Ma, "Time series forecasting of temperatures using sarima: An example from nanjing," p. 052024, 08 2018.

[11] R. Adhikari and R. Agrawal, *An Introductory Study on Time series Modeling and Forecasting*, 01 2013.

[12] ——, *An Introductory Study on Time series Modeling and Forecasting*, 01 2013.

[13] P. Zhang, E. Patuwo, and M. Hu, "Forecasting with artificial neural networks: The state of the art," *International Journal of Forecasting*, vol. 14, pp. 35–62, 03 1998.

[14] I. Khandelwal, R. Adhikari, and G. Verma, "Time series forecasting using hybrid arima and ann models based on dwt decomposition," vol. 48, 12 2014.

[15] G. Tunnicliffe Wilson, "Time series analysis: Forecasting and control,5th edition, by george e. p. box, gwilym m. jenkins, gregory c. reinsel and greta m. ljung, 2015. published by john wiley and sons inc., hoboken, new jersey, pp. 712. isbn: 978-1-118-67502-1," *Journal of Time Series Analysis*, vol. 37, pp. n/a–n/a, 03 2016.