National College of Ireland

Project Submission Sheet

Student Name:	Stamatios Karvounis		
Student ID:	x18197051		
Programme:	PGDDA_JAN23_O - Postgraduate Diploma in Science in Data Analytics	**Year:**	2023
Module:	Data Mining and Machine Learning II		
Lecturer:	Dr Abdul Razzaq		
Submission Due Date:	10/12/2023		
Project Title:	CA1 Data Mining and Machine Learning II		
Word Count:	6025		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Stamatios Karvounis

Date: 10/12/2023

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

Domain Application of Predictive Analytics

Predictive Model in Car Insurance Dataset

Your Name/Student Number	Course	Date
Stamatios Karvounis	Data Mining and Machine Learning II	10/12/2023

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

EVIDENCE OF AI USAGE

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[Place evidence here]

Additional Evidence:

[Place evidence here]

Time Series Predictive Models

Focus on Neural Networks

Stamatios Karvounis
School of Computing Information
National College of Ireland
line 4: Dublin, Ireland
x18197051@student.ncirl.ie

***Abstract*—This study investigates the effectiveness of advanced time series modeling techniques—Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, and Seasonal Autoregressive Integrated Moving Average (SARIMA)—in forecasting long-term maximum temperature (TMAX) trends using historical climate data. Employing the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, the research analyzes a comprehensive dataset of daily weather observations from 1980 to 2023, sourced from the National Oceanic and Atmospheric Administration. Each model was rigorously evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), supplemented by visualizations to compare predicted and actual data.**

Keywords— Time Series, Predictive Models, Neural Networks

I. Introduction

A. Motivation

The need for advanced time series analysis in the field of weather forecasting is both critical and evident, given the increasing complexity and variability of weather patterns in the modern era. Weather forecasting is a quintessential example of a domain where time series data is not only abundant but also carries immense significance due to its direct impact on numerous aspects of society, including agriculture, transportation, disaster management, and public safety.

Weather data exhibit strong seasonal patterns along with underlying trends. Advanced time series techniques like SARIMA (Seasonal Autoregressive Integrated Moving Average) allow for the explicit modeling of these patterns, thereby enhancing the accuracy of forecasts. Seasonality, whether it's the annual monsoon season or daily temperature cycles, can be effectively captured and predicted using these methods (Mishra and Desai, 2020).

Weather systems are inherently dynamic and non-linear, influenced by a multitude of factors ranging from local geographical conditions to global climatic changes. Traditional linear models often fall short in capturing these complexities. Advanced time series analysis, with its ability to model and predict such non-linear behaviors, becomes indispensable for accurate weather forecasting.

Weather conditions on a given day can be influenced by those of previous days or even weeks. This long-term dependency is something that traditional models might not account for adequately. Modern time series approaches like Long Short-Term Memory networks (LSTMs) in machine learning are designed to capture these dependencies, making them highly suitable for weather data (Zhao and Ghosh, 2020).

With the advent of IoT and advanced sensors, real-time weather data collection has become more feasible. This influx of continuous data necessitates models that can adapt and update their forecasts rapidly. Advanced time series models, especially those incorporating machine learning, are adept at learning from new data as it becomes available, thus providing up-to-date and reliable forecasts (Perera et al., 2018).

Weather forecasting involves multiple variables (like temperature, humidity, wind speed) that interact in complex ways. Advanced time series analysis allows for the integration of these multiple features, understanding their interactions, and how they jointly influence future weather patterns.

Accurate weather predictions are essential for risk mitigation in extreme weather events, resource management, and strategic planning across various sectors. Improved predictive capabilities through advanced time series analysis can lead to better preparedness and response strategies for extreme weather events, potentially saving lives and reducing economic losses (Zhao and Ghosh, 2020).

B. Research Question

This research aims to answer the following question: "How do advanced time series modeling techniques, specifically RNN, LSTM networks, and SARIMA models, compare in their effectiveness and accuracy for forecasting long-term maximum temperature (TMAX) trends using extensive historical climate data?"

This research question aims to explore and evaluate the efficacy of these sophisticated modeling approaches in the context of climate data analysis, particularly focusing on their predictive capabilities, strengths, and limitations in forecasting maximum temperature trends. The question also implies an underlying investigation into the computational challenges and the adaptation required in handling extensive seasonal patterns in time series data.

C. Structure of the Paper

The document is structured as follows: In the Related Work segment, the paper synthesizes key findings from prior research on deep learning and time series forecasting, critically assessing their relevance to the study at hand. The Data Mining Methodology section delineates the application of the CRISP-DM framework, elaborating on each phase from data understanding to deployment. The Data Preprocessing and Feature Engineering part outlines the procedures for data cleansing, addressing missing values, and feature engineering, such as extracting date-time features and computing rolling window statistics.

The Rationale for Model Choice discusses the selection of RNN, LSTM, and SARIMA models, focusing on their appropriateness for time series analysis in the context of weather forecasting. Experiment Setup details the division of data for training and testing and the development and training of RNN and LSTM models. The Evaluation/Results section encompasses the evaluation metrics utilized (MAE, RMSE,

MAPE) and presents the findings and visual comparisons for the SARIMA, RNN, and LSTM models.

Analysis of Patterns and Anomalies explores the patterns and irregularities unearthed during the research, with a particular emphasis on the challenges related to the seasonal periodicity of the SARIMA model. The Conclusions and Future Work segment summarizes the principal findings, discusses the study's limitations, proposes areas for future research, and underscores the paper's contribution to the fields of data mining and machine learning in time series analysis. Finally, the References section lists the scholarly works cited throughout the paper, formatted according to the requisite citation style

II. Data Variables Description

The dataset for this study was sourced from the National Oceanic and Atmospheric Administration's National Centers for Environmental Information. An order for the dataset was placed on December 4, 2023, requesting comprehensive daily weather observations. The requested data spanned a significant period, beginning on January 1, 1980, and extending to November 30, 2023. This extensive dataset, consisting of 16,040 daily observations, provides a robust basis for detailed climatic analysis.

Data procurement focused on Station USW00024121, with the specified unit of measurement being the standard format. The initial request encompassed a broader date range, starting from January 1, 1950, to December 31, 2009. However, the actual dataset received was tailored to the aforementioned specific period of interest.

The dataset included various data types essential for comprehensive weather analysis, such as ACSH, ADPT, ASLP, AWND, PRC, PRH, AVSN, SNOW, SNWD, TAVG, TMAX, TMIN, and TSUN. The inclusion of these specific data types ensures a multi-faceted approach to understanding and analyzing weather patterns over the long term. The delivery of this dataset via email facilitated immediate access, allowing for prompt commencement of data analysis and research activities.

The dataset is a valuable resource for model training in weather forecasting applications. Its comprehensive coverage of daily weather parameters over several years makes it an ideal candidate for time series analysis and forecasting models. Researchers and practitioners can leverage this dataset to develop and test predictive models, particularly focusing on Indian climate dynamics. The dataset's granularity and duration offer an opportunity to explore seasonal patterns, long-term trends, and day-to-day weather fluctuations, providing insights that could be critical for various practical applications, from urban planning to agricultural scheduling.

III. Related Work

The application of deep learning and time series forecasting methods, particularly SARIMA (Seasonal Autoregressive Integrated Moving Average) and LSTM (Long Short-Term Memory) models, has been widely explored in various studies. These works offer valuable insights into the capabilities and limitations of these methods in different contexts.

Chen et al. (2018) utilized SARIMA to forecast temperatures in Nanjing, demonstrating good forecasting accuracy and potential for future applications. The study showed that SARIMA could effectively predict mean temperature over 36 months using 35 years of data. Their approach highlighted the importance of testing forecasting accuracy through comparison with observational values to avoid underfitting or overfitting. The study's strength lies in its detailed statistical analysis and practical testing, while its limitation is in the specific regional focus, which may not generalize to other locations or datasets.

Dubey et al. (2021) compared SARIMA and LSTM for forecasting energy consumption. The study found LSTM more effective than SARIMA, highlighting the importance of considering various weather features in energy consumption prediction. They demonstrated the high correlation of energy consumption with factors like humidity and temperature. This study's comprehensive approach to feature selection and its application to smart grid data is a significant strength. However, its focus on energy consumption limits its direct applicability to broader climate data forecasting.

Dimri et al. (2020) used SARIMA to analyze climate variables, specifically temperature and precipitation, in the Bhagirathi river basin. Their findings underscored SARIMA's effectiveness in making long-term forecasts (up to 20 years). The study's strength lies in its long-term analysis, but it also noted over-predictions in extreme events, highlighting a potential weakness in handling anomalies.

These studies collectively demonstrate the effectiveness of SARIMA and LSTM in time series forecasting, with each method having its strengths in different contexts. SARIMA shows robust performance in linear, seasonal patterns, while LSTM excels in capturing long-term dependencies and non-linear relationships.

The literature on the application of deep learning, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models, for time series forecasting, especially in weather prediction, provides comprehensive insights and diverse methodologies.

Karevan & Suykens (2020) introduced Transudative LSTM (T-LSTM) for weather forecasting, emphasizing its capability to capture long-term dependencies and focus on local information near test points. T-LSTM's use of weighted quadratic cost functions based on cosine similarity between training samples and test points shows improved performance in weather prediction. This study is foundational in demonstrating LSTM's effectiveness in forecasting and its adaptability to transudative learning, though the focus on quadratic cost functions might limit its applicability in more complex, non-linear scenarios.

Singh et al. (2019) compared SVM, ANN, and RNN for weather forecasting, highlighting RNN's superior performance in time series prediction. This research validates the effectiveness of RNNs in handling sequential weather data. However, its limitation lies in the scope of parameters considered for forecasting, which may not capture the complete complexity of weather patterns.

Salman et al. (2015) investigated deep learning techniques for weather forecasting, comparing RNN, CRBM, and CN models. The study underlines the potential of deep learning models in identifying complex patterns in weather data, but the comparison lacks a detailed exploration of each model's specific strengths and weaknesses in different weather forecasting contexts.

Balluff et al. (2020) explored meteorological data forecasting using RNNs. Their work on forecasting wind speed using RNNs contributes to the understanding of the practical application of deep learning in energy-related meteorology. While their findings affirm RNN's utility in forecasting, the study indicates a need for improved robustness and accuracy.

For the current research project, these studies serve as a foundation in understanding the applicability and potential of RNN and LSTM models for long-term climate data analysis. They demonstrate the effectiveness of these models in capturing temporal dependencies, an essential aspect for accurately forecasting maximum temperature trends. The methodologies and findings from these studies can guide the project's approach to model selection, implementation, and evaluation, ensuring a comprehensive and well-informed analysis.

IV. Data Mining Methodology

A. *CRISP-DM*

In this project focusing on time series analysis of climate data, the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework was methodically applied. This framework, renowned for its structured approach to data mining projects, encompasses six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Here is how each phase was implemented in the project:

Business Understanding

The initial phase involved defining the objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition. For this project, the goal was to analyze and forecast climate variables, particularly focusing on TMAX. The end objective was to gain insights into climate trends and patterns, which could be valuable for various applications, including urban planning and climate research.

Data Understanding

This phase started with data collection, where daily climate data was gathered. The dataset included variables such as max temperature, min temperature, snow, precipitation, sun indicators. Initial data exploration involved assessing the quality of the data, identifying the key variables, and understanding their temporal nature. This was crucial in guiding the subsequent data preparation and modeling steps.

Data Preparation

Data cleaning and preprocessing were undertaken, where the datasets were first concatenated to form a comprehensive set. The data was then checked and cleaned for missing values, duplicates, and outliers. Feature engineering was conducted to enhance the dataset with additional time-based and statistical features, crucial for time series analysis.

Modeling

In the modeling phase, the focus was on applying the SARIMA, RNN and LSTM models, suitable for time series forecasting. This involved checking the stationarity of the time series, determining the parameters for the SARIMA model, and then fitting the model to the data. Additional analyses like decomposition of the time series and autocorrelation studies were performed to better understand the data and refine the model.

Evaluation

The model's performance was evaluated based on its ability to accurately forecast future values. This involved comparing the model's predictions against actual data, assessing the fit of the model, and ensuring that the residuals were behaving appropriately..

B. *Data Processing, Feature Selection and Engineering*

The dataset, encompassing 16,040 daily observations from January 1, 1980, to November 30, 2023, was initially subjected to a thorough examination to ascertain its structure and identify any missing values. It comprised various meteorological parameters, including precipitation (PRCP), snowfall (SNOW), snow depth (SNWD), maximum temperature (TMAX), minimum temperature (TMIN), average wind speed (AWND), evaporation (EVAP), and possible sunshine (PSUN).

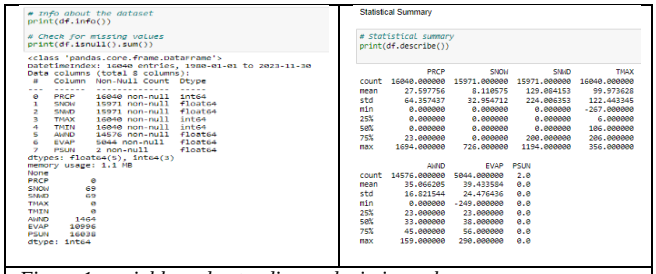


Figure 1: variable understanding and missing values

A pivotal aspect of the preprocessing involved addressing missing data, a common challenge in such extensive datasets. Specifically, missing values in SNOW, SNWD, and AWND were filled using the forward fill method. However, due to the significant amount of missing data in EVAP and PSUN, these columns were excluded from the analysis to maintain data integrity.

Further, the preprocessing phase also included the detection and capping of outliers. This step was crucial to mitigate the impact of extreme values on the analysis. The dataset's numeric columns, such as PRCP, SNWD, TMAX, TMIN, and AWND, were adjusted based on the interquartile range (IQR), effectively addressing potential skewness caused by outliers.

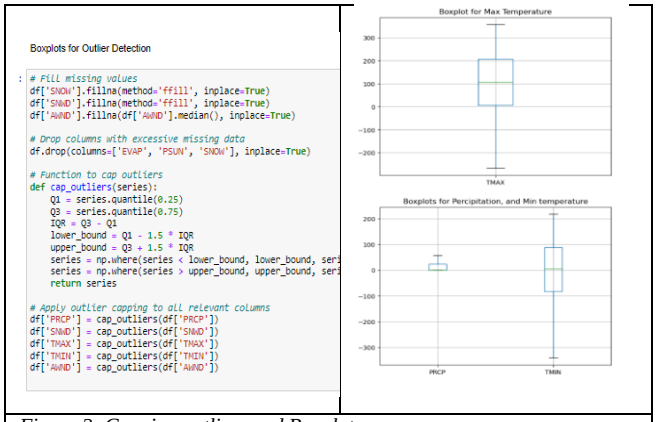


Figure 2: Capping outliers and Boxplots

For feature engineering, the study capitalized on the rich temporal nature of the dataset. Essential date-time features

like year, month, day, day of the week, day of the year, quarter, and week of the year were extracted from the datetime index. These features were instrumental in capturing the temporal dynamics and seasonal patterns inherent in the data.

To further enhance the dataset, rolling window features, such as the rolling mean and standard deviation of TMAX, were computed over a 7-day period. These features provided insights into the short-term trends and variations in temperature. Additionally, lag features for TMAX, specifically one-day and two-day lags, were introduced to incorporate the influence of preceding days' temperatures on subsequent readings.

Date-Time Features

```
# Extracting date-related features
df['year'] = df.index.year
df['month'] = df.index.month
df['day'] = df.index.day
df['dayofweek'] = df.index.dayofweek
df['dayofyear'] = df.index.dayofyear
df['quarter'] = df.index.quarter
df['weekofyear'] = df.index.isocalendar().week
```

Rolling Window Features

```
# Rolling window features
window_size = 7  # 7 days for a weekly pattern

df['rolling_TMAX'] = df['TMAX'].rolling(window=window_size).mean()
df['rolling_std_TMAX'] = df['TMAX'].rolling(window=window_size).std()
```

Lag Features

```
# Lag features
df['lag_1_temp'] = df['TMAX'].shift(1)
df['lag_2_temp'] = df['TMAX'].shift(2)
```

Figure 3: date-time features, rolling window, lag features

The exploratory data analysis (EDA) phase of the study was comprehensive and multifaceted. It involved a basic data overview and visualization of key variables like TMAX, PRCP, SNOW, and TMIN. The visualizations shed light on the trends, seasonality, and variations within the dataset. Seasonality and trend analysis were conducted through seasonal decomposition of the TMAX time series, which facilitated the separation of the series into trend, seasonal, and residual components.

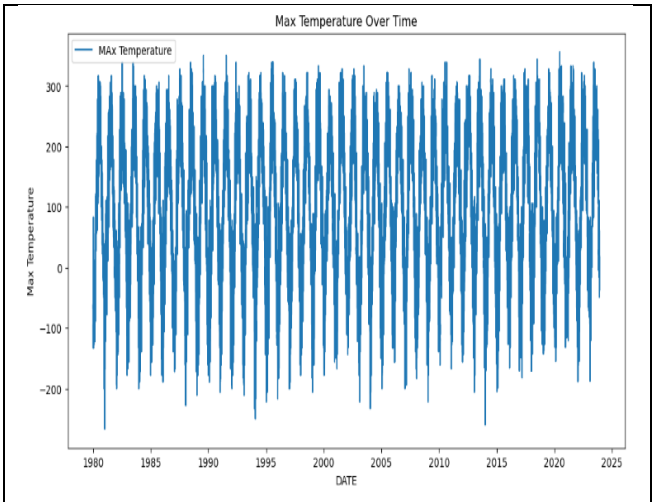


Figure 4: Max temperature over time

The distribution of TMAX was examined using histograms and Kernel Density Estimation plots, offering a detailed perspective on the variable's spread and density. Rolling statistics, including a 30-day rolling mean and standard deviation for TMAX, were calculated and visualized to highlight the moving average trends and variability over time.

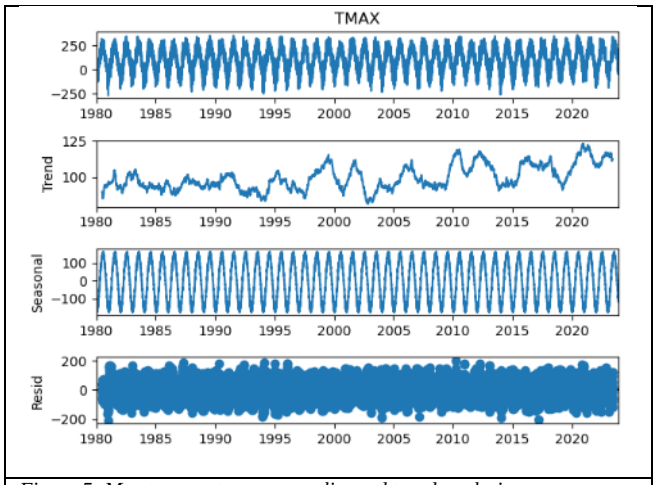


Figure 5: Max temperature seasonality and trend analysis

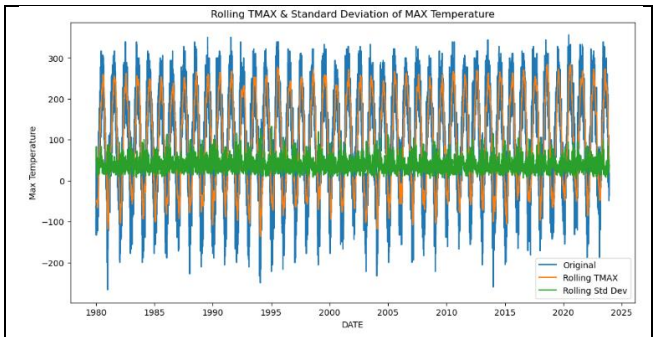
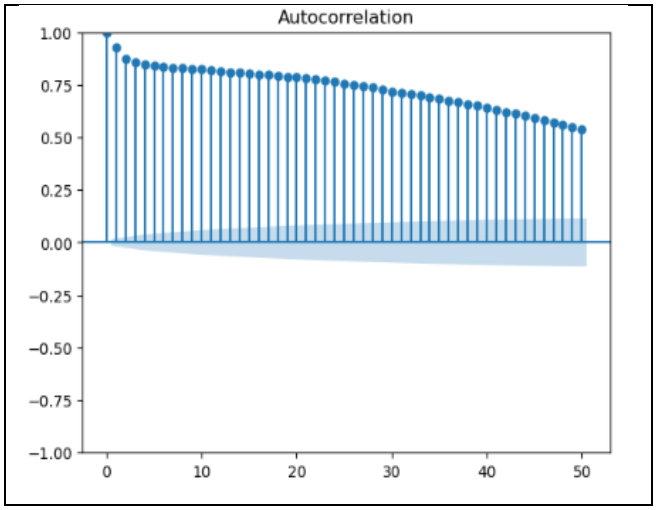


Figure 6: Max temperature rolling statistics

Additionally, autocorrelation and partial autocorrelation analyses were performed on TMAX to discern the degree of correlation at different time lags. Lastly, a correlation matrix was constructed and visualized through a heatmap, elucidating the interrelationships among the various weather variables.



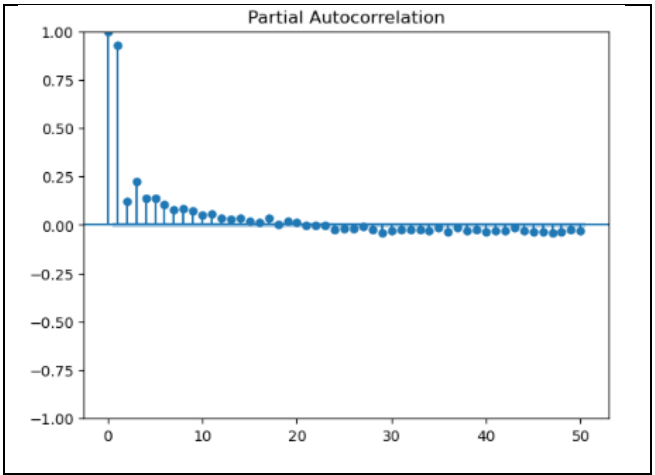


Figure 7: Autocorrelation and Partial Autocorrelation

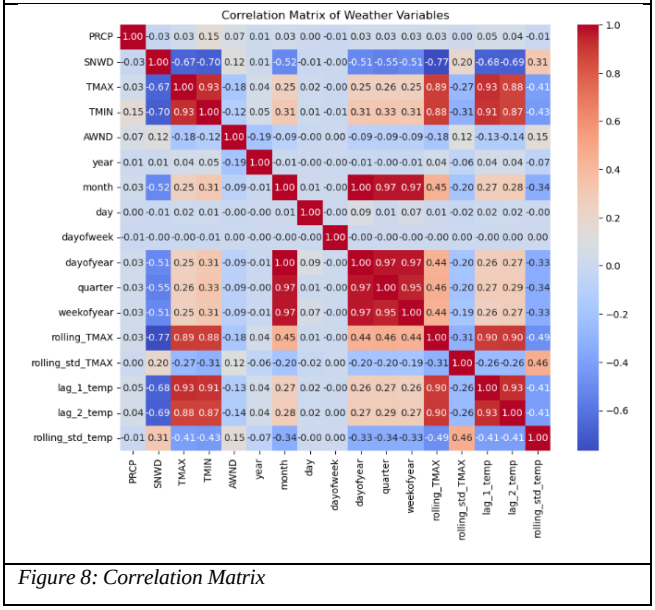


Figure 8: Correlation Matrix

Through these meticulous preprocessing and exploratory steps, the dataset was effectively transformed and enriched, laying a solid foundation for subsequent analytical modeling. The employment of both statistical and visual techniques in the EDA provided an in-depth understanding of the dataset's characteristics, underlying patterns, and potential predictive relationships.

C. Rationale for choosing SARIMA, RNN, LSTM

The rationale for selecting SARIMA (Seasonal Autoregressive Integrated Moving Average), RNN (Recurrent Neural Network), and LSTM (Long Short-Term Memory) models for the study lies in their specific strengths and capabilities in handling time series data, particularly in forecasting future values based on historical data. Each of these models brings unique advantages to the table, making them well-suited for time series analysis in the context of climate data.

SARIMA

- Seasonality and Non-Stationarity: SARIMA is an extension of the ARIMA model that specifically accounts for seasonality, making it highly suitable for datasets with clear seasonal patterns, like climate data.
- Interpretable Parameters: SARIMA models have interpretable parameters that can provide insights into the data's underlying patterns, such as trends and seasonality.
- Effectiveness with Linear Relationships: SARIMA works well with time series data where relationships are more linear and can be described using differences and moving averages.

RNN

- Handling Sequential Data: RNNs are designed to work with sequential data. They can capture temporal dependencies and patterns in time series data, which is essential for accurately predicting future climate variables.
- Memory of Past Information: RNNs have a 'memory' that captures information about what has been calculated so far, making them advantageous for datasets where past information is a strong predictor of future events.

LSTM

- Long-Term Dependencies: LSTMs are a special kind of RNN capable of learning long-term dependencies. This is particularly useful in climate data, where recent conditions as well as conditions from the more distant past can influence future climate patterns.
- Overcoming Vanishing Gradient Problem: LSTMs are designed to avoid the long-term dependency problem, making them more effective than traditional RNNs for many time series datasets.
- Flexibility with Non-Linear Patterns: Unlike SARIMA, LSTMs can capture non-linear relationships, which can be present in complex time series data like weather patterns.

In summary, the choice of these models was driven by their compatibility with the characteristics of time series data, particularly in the context of climate forecasting. SARIMA offers a robust approach for linear, seasonal patterns, while RNN and LSTM provide the capability to model more complex, non-linear relationships and long-term dependencies in the data. Combining these approaches in the study provided a comprehensive and versatile analysis, harnessing the strengths of each method to address different aspects of the time series data.

D. Setup of Experiments

In this study, a methodical approach was adopted to set up the experiments, encompassing data partitioning for training, validation, and testing, along with the deployment of two distinct neural network models: Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks.

Data Partitioning

The dataset, sourced from the National Oceanic and Atmospheric Administration's National Centers for Environmental Information, included daily meteorological readings from January 1, 1980, to November 30, 2023. For the purpose of this study, the dataset was partitioned into training and testing sets to evaluate the models' performance and generalization capabilities. The partitioning was executed with a temporal split, using data up to January 1, 2023, for training, and data post this date for testing. This approach ensured that the models were trained and evaluated on non-

overlapping, sequential data segments, maintaining the temporal integrity essential for time series analysis.

Model Development and Training

Two advanced neural network architectures were employed for the experiment:

1. Recurrent Neural Network (RNN): The RNN model was designed to capture temporal dependencies in the data. It consisted of a SimpleRNN layer with 50 units, followed by a Dense layer for output prediction. The input data were reshaped to match the [samples, time steps, features] format required by RNNs.

2. Long Short-Term Memory (LSTM) Network: The LSTM model, known for its efficacy in handling long-term dependencies, was structured with an LSTM layer comprising 50 units, succeeded by a Dense output layer. Similar to the RNN, the input data for the LSTM were also reshaped appropriately.

Both models were compiled with a mean squared error loss function and optimized using the Adam optimizer. The training process for each model spanned 10 epochs with a batch size of 1, ensuring thorough learning from the sequential data.

Evaluation and Comparison of Models

The performance of the RNN and LSTM models was meticulously evaluated using three key metrics:

- Root Mean Squared Error (RMSE): Quantifying the square root of the average squared differences between predicted and actual values.
- Mean Absolute Error (MAE): Measuring the average magnitude of errors in predictions.
- Mean Absolute Percentage Error (MAPE): Expressing the error as a percentage of actual values, offering a relative error measurement.

These metrics were computed for both the training and testing sets, enabling a comprehensive assessment of each model's predictive accuracy and generalization ability.

Visualization

To visually interpret the models' performance, time series plots were created, juxtaposing the actual TMAX values against the predicted values for both training and testing phases. These visualizations provided an intuitive understanding of the models' predictive behavior over time.

Evaluation Metrics

The models were evaluated using metrics appropriate for time series forecasting, such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics provided insights into the accuracy and reliability of the models in forecasting future climate variables.

This structured experimental setup, with clear data partitioning and systematic model evaluation, was critical in ensuring the robustness and validity of the forecasting models developed in the study.

V. Evaluation-Results

In the subsequent phase of the project, a detailed analysis was conducted to evaluate the performance of the Seasonal Autoregressive Integrated Moving Average (SARIMA), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) models. The evaluation centered on three primary metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). Additionally, visualizations were generated to compare the predicted data against actual observations, providing a comprehensive and intuitive assessment of model performance.

Evaluation Metrics

1. Mean Absolute Error (MAE): This metric measures the average magnitude of errors in the predictions, without considering their direction. It is a straightforward representation of average error and was chosen for its simplicity and interpretability.

2. Root Mean Squared Error (RMSE): RMSE calculates the square root of the average squared differences between predicted and actual values. This metric was selected due to its sensitivity to large errors, making it particularly useful in identifying and penalizing significant prediction discrepancies.

3. Mean Absolute Percentage Error (MAPE): MAPE expresses the error as a percentage of actual values, offering a relative measure of error. It was included to provide a normalized perspective, enabling the comparison of model performance across different scales or units of measurement.

Model Evaluations

- SARIMA: The SARIMA model exhibited a MAE of 135.58 and an RMSE of 163.50. However, the MAPE was anomalously high, which indicated a need for further investigation or model refinement.
- RNN: The RNN model achieved a train RMSE of 46.59, MAE of 35.56, and MAPE of 66.98%. For the test set, the RMSE was 45.29, MAE was 35.03, and MAPE was 76.11%, demonstrating the model's consistent performance across both training and testing phases.
- LSTM: The LSTM model presented a train RMSE of 46.89, MAE of 35.51, and MAPE of 65.93%. The test scores were slightly better, with an RMSE of 45.59, MAE of 34.97, and MAPE of 74.91%.

Visualizations

To augment the quantitative analysis, visualizations were crafted to compare the predicted data from each model against the actual observations. These plots provided a graphical representation of the models' predictive accuracy over time, enabling a more intuitive understanding of model performance. The visualizations were particularly useful in identifying patterns, trends, and anomalies in the predictions, thereby offering a more holistic view of the models' capabilities and areas for improvement.

The use of MAE, RMSE, and MAPE as evaluation metrics offered a comprehensive assessment of the models' predictive accuracy and error characteristics. The visual comparisons between predicted and actual data served as a valuable tool for validating the quantitative findings and gaining deeper insights into the models' behavior. The combination of these

evaluation strategies facilitated a robust analysis of the SARIMA, RNN, and LSTM models, providing a solid foundation for further refinement and application in the field of meteorological data analysis.

SARIMA Model Analysis

The SARIMA model, known for its efficacy in capturing seasonal patterns in time series data, was the first model employed. It yielded a Mean Absolute Error (MAE) of 135.58 and a Root Mean Squared Error (RMSE) of 163.50. However, the model's Mean Absolute Percentage Error (MAPE) was exceptionally high, indicating potential issues in the model's predictive accuracy or an anomaly in the data that requires further investigation.

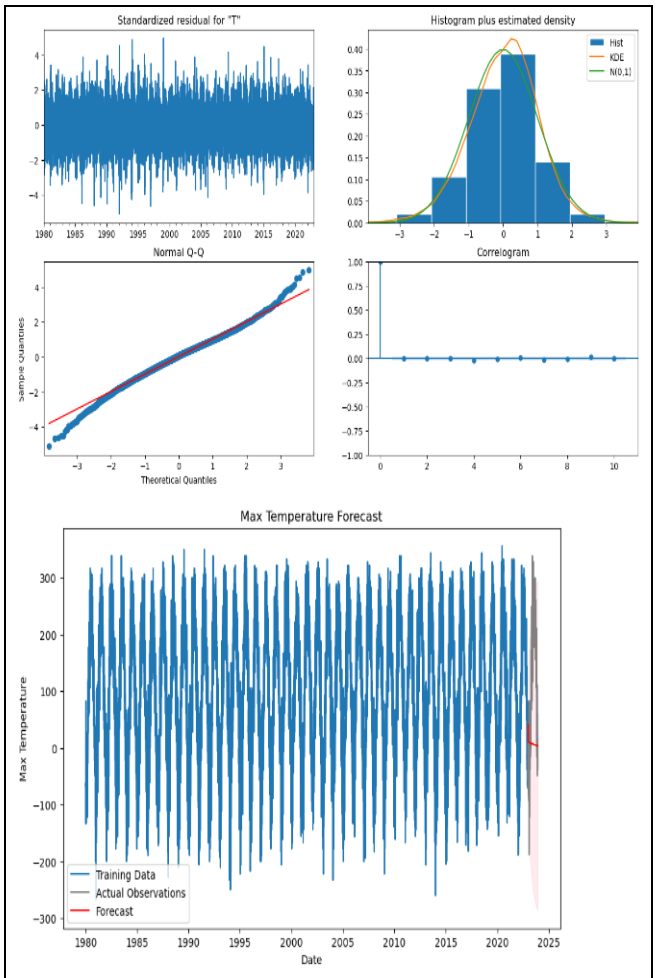


Figure 9: SARIMA forecasting

RNN Model Analysis

The RNN model, designed to leverage the temporal dynamics in sequential data, demonstrated notable consistency in its performance. For the training set, it achieved an RMSE of 46.59, an MAE of 35.56, and a MAPE of 66.98%. In the testing phase, the model maintained its performance with an RMSE of 45.29, an MAE of 35.03, and a MAPE of 76.11%. These results underscore the model's robustness in handling the dataset and its generalizability across different data segments.

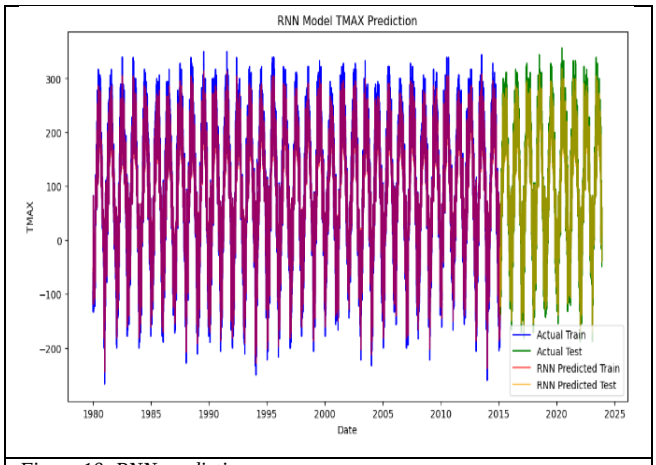


Figure 10: RNN prediction

LSTM Model Analysis

Finally, the LSTM model, renowned for its ability to capture long-term dependencies in time series data, presented slightly varied yet competitive results. The LSTM model's training scores included an RMSE of 46.89, an MAE of 35.51, and a MAPE of 65.93%. On the testing set, it recorded an RMSE of 45.59, an MAE of 34.97, and a MAPE of 74.91%. These results indicate that the LSTM model was marginally more effective in predicting the test data, suggesting its superior capability in understanding and leveraging the underlying patterns in the temperature data.

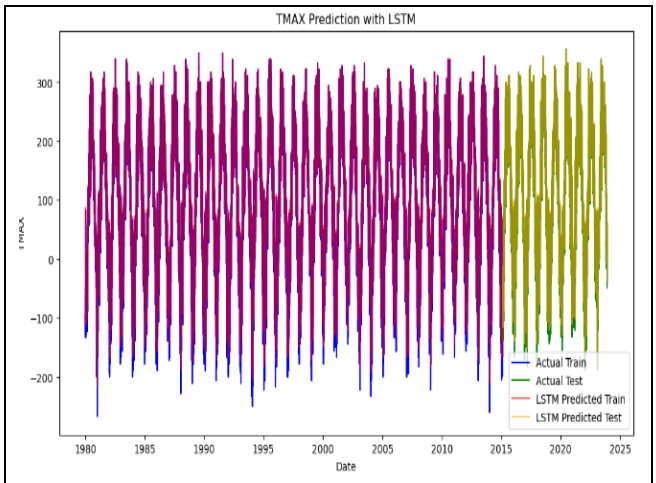


Figure 11: LSTM prediction

A. *Obstacles*

D In the process of modeling the daily weather data using the SARIMA (Seasonal Autoregressive Integrated Moving Average) model, the treatment of seasonality in the dataset was a pivotal factor. The dataset, comprising daily observations, naturally suggested the use of a yearly cycle (m=365) to capture the seasonal variations inherent in such data. However, this approach presented significant computational challenges.

The computational demands of processing a full year's seasonality in a daily dataset are considerable. When the seasonal periodicity was set to m=365, it resulted in formidable computational requirements, particularly in terms of memory usage. This issue arose from the SARIMA model's

complexity in dealing with the extensive seasonal cycles over a prolonged period. Consequently, the computing resources available for this study were insufficient to manage such an extensive degree of seasonality.

To circumvent this computational hurdle, a strategic decision was made to adjust the seasonal periodicity to m=7, thus focusing on a weekly seasonal pattern. This adjustment substantially alleviated the computational load, enabling the SARIMA model to be fitted without memory constraints. However, it is essential to recognize that this modification introduces a limitation in capturing the full spectrum of the yearly seasonal trends in the weather data.

The SARIMA model with a seasonal periodicity of m=7 primarily captures weekly patterns and may not fully represent the more extended annual seasonal dynamics. Therefore, the results derived from this model should be interpreted with an understanding of this constraint. The model's potential inability to fully encapsulate the annual seasonality could impact the accuracy and generalizability of its forecasts. In future research endeavors where more powerful computational resources are available, a reevaluation using the ideal seasonal periodicity of m=365 would be beneficial to more accurately model the yearly seasonal patterns in daily weather data.

VI. Conclusion-Future Work

A The study focused on a comprehensive dataset encompassing 16,040 daily observations from 1980 to 2023, provided by the National Oceanic and Atmospheric Administration's National Centers for Environmental Information. The key findings are as follows:

SARIMA Model: Exhibited moderate predictive accuracy with an MAE of 135.58 and RMSE of 163.50. However, an anomalously high MAPE suggests the need for further model refinement or data investigation.

RNN Model: Showed consistent performance across training and testing datasets, with RMSE scores of 46.59 and 45.29, respectively. This model demonstrated robust generalizability and effective handling of temporal dynamics in the data.

LSTM Model: Marginally outperformed the RNN in testing phases, indicating its enhanced ability to capture and leverage long-term dependencies in the dataset.

Visualizations provided clear insights into each model's predictive capacity, revealing close alignments between the RNN and LSTM predictions with the actual data trends.

Limitations and Future Work

The study encountered a significant limitation in the SARIMA model's application due to computational constraints. The model's inability to process the optimal yearly seasonal periodicity (m=365) led to a compromise with a weekly cycle (m=7). This adjustment, while necessary, potentially reduced the model's capability to fully capture the annual seasonal trends, impacting the forecast accuracy.

Future work should focus on addressing these computational challenges, perhaps by leveraging more robust computing resources or exploring alternative modeling techniques that can efficiently handle extensive seasonal data. Reevaluating the dataset with the SARIMA model using the ideal m=365 setting could yield more accurate and reliable forecasts.

Further research might also explore hybrid models that combine the strengths of SARIMA, RNN, and LSTM, potentially leading to improved predictive performance. Additionally, incorporating external factors such as climate change indicators could enhance the models' comprehensiveness.

Contribution to the Field

This paper contributes to the field of data mining and machine learning, particularly in time series analysis, by demonstrating the application of SARIMA, RNN, and LSTM models in meteorological data forecasting. The comparative analysis of these models provides valuable insights into their respective strengths and limitations in handling complex time series data. The research underscores the importance of considering computational constraints and seasonal periodicity in model selection and highlights the potential of visualizations in enhancing the interpretability of model outputs.

The study's findings and methodologies offer a foundation for future research endeavors in time series forecasting, encouraging the exploration of advanced modeling techniques and hybrid approaches in the domain of environmental data analysis. The research also sets a precedent for addressing computational limitations creatively, paving the way for more effective and efficient modeling strategies in the field.

References

[1] Abayomi-Alli, A., Odusami, M. O., Abayomi-Alli, O., Misra, S., & Ibeh, G. F. (2019, July). Long short-term memory model for time series prediction and forecast of solar radiation and other weather parameters. In 2019 19th international conference on computational science and its applications (ICCSA) (pp. 82-92). IEEE.

[2] Balluff, S., Bendfeld, J., & Krauter, S. (2020). Meteorological data forecast using RNN. In Deep Learning and Neural Networks: Concepts, Methodologies, Tools, and Applications (pp. 905-920). IGI Global.

[3] Chen, P., Niu, A., Liu, D., Jiang, W., & Ma, B. (2018, August). Time series forecasting of temperatures using SARIMA: An example from Nanjing. In IOP Conference Series: Materials Science and Engineering (Vol. 394, p. 052024). IOP Publishing.

[4] Dabral, P. P., & Murry, M. Z. (2017). Modelling and forecasting of rainfall time series using SARIMA. Environmental Processes, 4(2), 399-419.

[5] Dimri, T., Ahmad, S., & Sharif, M. (2020). Time series analysis of climate variables using seasonal ARIMA approach. Journal of Earth System Science, 129, 1-16.

[6] Dubey, A. K., Kumar, A., García-Díaz, V., Sharma, A. K., & Kanhaiya, K. (2021). Study and analysis of SARIMA and LSTM in forecasting time series data. Sustainable Energy Technologies and Assessments, 47, 101474.

[7] Farsi, M., Hosahalli, D., Manjunatha, B. R., Gad, I., Atlam, E. S., Ahmed, A., ... & Ghoneim, O. A. (2021). Parallel genetic algorithms for optimizing the SARIMA model for better forecasting of the NCDC weather data. Alexandria Engineering Journal, 60(1), 1299-1316.

[8] Hewage, P., Behera, A., Trovati, M., Pereira, E., Ghahremani, M., Palmieri, F., & Liu, Y. (2020). Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. Soft Computing, 24, 16453-16482.

[9] Karevan, Z., & Suykens, J. A. (2020). Transductive LSTM for time-series prediction: An application to weather forecasting. Neural Networks, 125, 1-9.

[10] Mishra, A. K., & Desai, V. R. (2020). Big data analytics for weather prediction. International Journal of Information Technology, 12(3), 731-738.

[11] Perera, C., Qin, Y., Estrella, J. C., Reiff-Marganiec, S., & Vasilakos, A. V. (2018). Fog computing for sustainable smart cities: A survey. ACM Computing Surveys (CSUR), 50(3), 1-43.

[12] Salman, A. G., Kanigoro, B., & Heryadi, Y. (2015, October). Weather forecasting using deep learning techniques. In 2015 international conference on advanced computer science and information systems (ICACSIS) (pp. 281-285). Ieee.

[13] Singh, S., Kaushik, M., Gupta, A., & Malviya, A. K. (2019, March). Weather forecasting using machine learning techniques. In Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE).

[14] Zhao, N., & Ghosh, S. (2020). Long-term temporal CNN for daily-to-decadal climate extremes prediction: Case study with Barents and Kara sea ice. Atmosphere, 11(2).

[15] Zhao, N., & Ghosh, S. (2020). Long-term temporal CNN for daily-to-decadal climate extremes prediction: Case study with Barents and Kara sea ice. Atmosphere, 11(2), 131.