Weather Forecasting using Time Series Analysis

Aditya Tanna (202103023),* Rushi Vaghela (202103028),† and Jay Rathod (202103042)‡
Dhirubhai Ambani Institute of Information & Communication Technology,
Gandhinagar, Gujarat 382007, India
SC475, Time Series Analysis

We live in a world where Global warming is on the rise and weather is becoming unpredictable day-by-day. So, weather forecasting has become very important in today's world .Urban cities, particularly in cities like Los Angeles, are significantly influenced by varying humidity and temperature levels, impacting everything from public health to urban infrastructure. This project delves into the forecasting of temperature and humidity levels using a comprehensive dataset that spans several years, capturing daily temperature humidity measurements. We have used various Time series techniques that we have studied to accurately make predictions and help in making day-to-day life decisions dependent on environmental factors more easily.

Our first task was to find the dataset that has all the aspects that will help us in our analysis. We picked this dataset from Kaggle [1] - which has 5 years of data on various weather metrics - temperature, humidity, wind speed etc. We found this sufficient and picked this dataset up. Data cleaning and preparation was done to ensure accuracy and reliability on the data that was to be used. Through exploratory data analysis, the study identifies significant patterns and anomalies in the temperature humidity trends.

We then looked at the Time Series aspects - Stationarity,Trend and Seasonality to make accurate judgements on the patterns that were being followed. We then performed stationarity checks and autocorrelation analyses, which are essential for effective time series forecasting. The SARIMAX model was then used to try and fit into our historical data, incorporating both seasonal and non-seasonal elements, to forecast future humidity levels. The forecasting model aims to provide insights into future humidity trends and their potential impacts on urban living.

The primary goal of this study is to offer a robust framework that helps people in city make their life decisions and daily plans based on predicted temperature and humidity levels.

I. INTRODUCTION

In recent years, the study of atmospheric conditions through statistical modeling has gained paramount importance due to its wide-ranging applications in agriculture, urban planning, and public health. This project focuses on the analysis and forecasting of temperature humidity levels in urban cities using a comprehensive dataset that captures daily temperature humidity trends over several years. The motivation behind selecting this particular dataset stems from the significant impact that temperature humidity has on climate comfort, environmental quality, and energy consumption patterns in urban settings. Our purpose remains the same, to create a framework that help the people in making informed decisions with ease based on the predictions that our model gives.

We have used the concepts taught in the course to do the following

- **Model Historical Data:** Fit a SARIMAX model to historical temperature and humidity data to capture both non-seasonal and seasonal dynamics.
- **Forecast Future Trends:** Use the model to forecast temperature and humidity for the next three years, providing valuable insights to the people.
- **Analyze Impact:** Understand the impact of various external factors on temperature and humidity levels, which could include urban development patterns, environmental policies, and climatic shifts.

At the end of the day, our aim is to help the community by providing this framework to the people that help them in their daily lives - our contribution to the community.

II. DATASET ANALYSIS

We have taken a weather dataset that we found online, which has different weather metrics (such as temperature, humidity, pressure, rainfall etc.) for a variety of cities (majorly in NA). We have **hourly data** from October 2012 to November 2017. For our analysis, we have taken two of the most important metrics for weather forecasting, i.e. **Temperature and Humidity**.

*Electronic address: 202103023@daiict.ac.in
†Electronic address: 202103028@daiict.ac.in
‡Electronic address: 202103042@daiict.ac.in

Out of all the cities, we have taken three cities, **Los Angeles, Miami and Jerusalem**. These three are selected for the following reasons.

- We have taken **Los Angeles and Miami** because both are the two American cities on different coasts, i.e. Los Angeles being West and Miami being East. This is done so we can compare and contrast the temperature and humidity variations based on the geographical locations of the two cities and make precise judgements on what are the factors affecting the metrics.
- Next, the reason we chose **Jerusalem** is because of the fact that it has a very different topology compared to the two cities. Another factor that could differently contribute is the **Elevation factor**. Jerusalem's inland location at a higher elevation compared to Los Angeles and Miami contributes to its wider temperature ranges and generally lower humidity levels, that we have analyzed.

A. The Temperature Data

We will now look at the Temperature data that we have in our temperature.csv file. We looked at the data for our three selected cities.

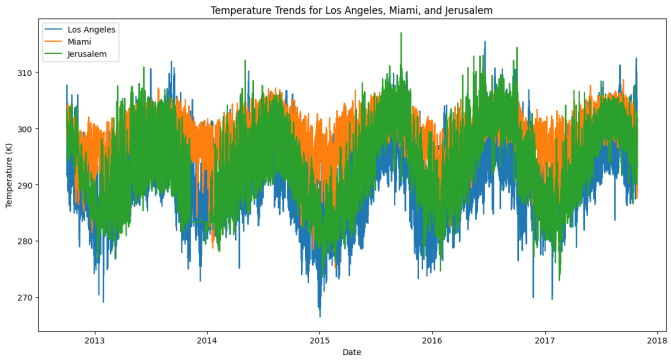

FIG. 1: Temperature Data for the three cities

The graph ((see 1) shows the Temperature trend across 5 years for the three cities, and there are a few inferences to be made

- **Los Angeles**: The temperature exhibits a stable pattern with moderate fluctuations throughout the year. There is a clear seasonal variation with peaks typically in the mid-year, indicating warmer summer months.
- **Miami**: Miami shows a more consistent temperature with less variation compared to Los Angeles and Jerusalem. The temperature remains relatively high throughout the year, highlighting its tropical climate due to its geographical positioning.
- **Jerusalem**: Here we notice more significant fluctuations and a distinct seasonal pattern, with temperatures peaking in summer and dropping considerably in winter. This indicates a Mediterranean climate with hot summers and cooler winters.

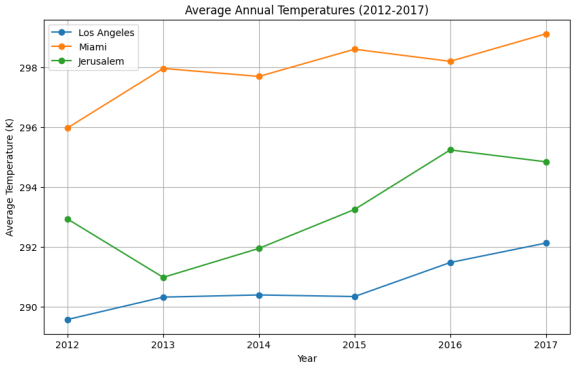

FIG. 2: Average Annual Temperature Comparison

Next, we looked at how the temperature was changing overtime in the three cities.

The graph (see 2) is the average annual temperature comparison across all the 3 cities. We see a gradual increase in the average temperature. This is due to the **Urban Heat Island Effect**. This means cities tend to become warmer than their rural surroundings. The possible contributing factors are increased energy usage (such as air conditioning), vehicle emissions, and industrial activity contribute to this effect.

B. The Humidity Data

Here, we will look at the humidity aspect for the three selected cities, and what factors are affecting such values

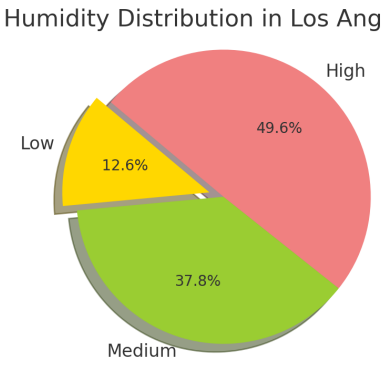

FIG. 3: Humidity Distribution in Los Angeles

For **Los Angeles**, the humidity trend is relatively stable compared to temperature, which tell us that **changes**

in temperature are not heavily influenced by large shifts in humidity levels. The coastal location helps in maintaining a stable humidity level, preventing extreme dryness despite the warm summers. (see 3)

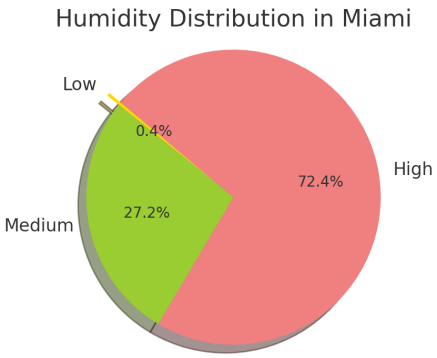

FIG. 4: Humidity Distribution in Miami

In **Miami**, We see high humidity throughout the year with very high levels during summer, reflecting the influence of its tropical setting and proximity to the sea, which contributes to a moist, warm climate. (see 4)

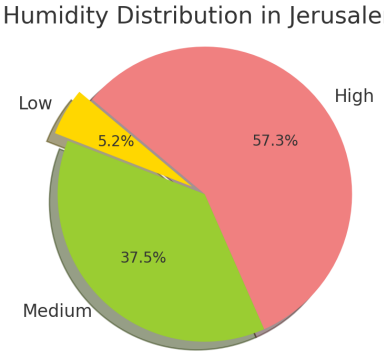

FIG. 5: Humidity Distribution in Los Angeles

Finally, looking at **Jerusalem**, there is low humidity compared to the other coastal cities, with peaks during the colder months due to seasonal rainfall. The city's higher elevation and inland position account for these characteristics. (see 5)

So, we analyzed the data and saw how we can compare the various city weather statistics and what might be the factors that affect these metrics (eg. Geographical location).

C. Rolling Averages

Our first question here would be, why do we even find rolling averages and how do they help ? The answer is simple! Rolling trends, such as a 30-day rolling average, are used in time series analysis to smooth out short-term fluctuations and highlight longer-term trends in data and provide a cleaner, more understandable visual representation of the trends. There might be some underlying trends hidden that we wouldn't see normally so we do rolling trend analysis to acquire those trends. We have done it for our selected data.

We looked at both Temperature and Humidity data in one graph for all the three cities and we try and plot them both together on a time scale to check their dependence on one another.

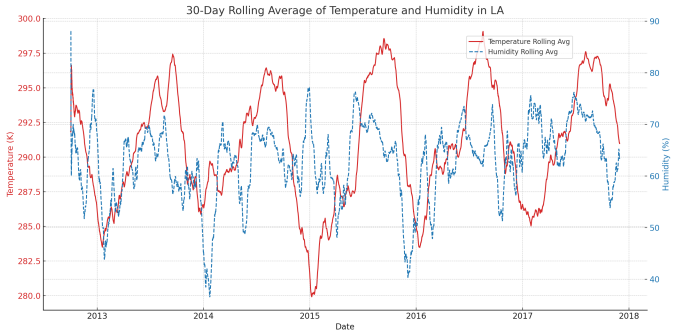

FIG. 6: 30-Day Rolling Average of Temperature and Humidity in LA

The above is the **30 Day Rolling Average for Los Angeles**. We can clearly see the seasonal pattern of temperature and humidity with **highs during summer and lows during winters.** This is the temperature pattern.Humidity, inversely, tends to be higher in the cooler months and lower when the temperatures peak, indicating an inverse relationship between temperature and humidity for Los Angeles. (see 6)

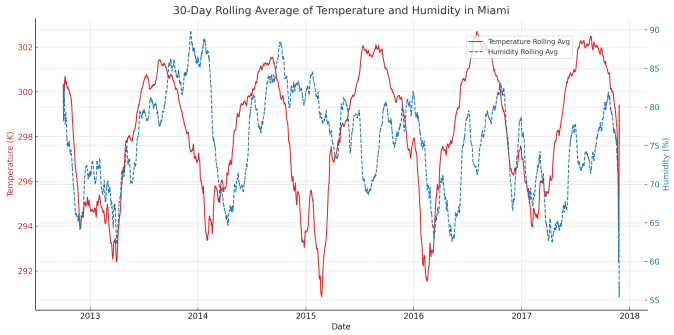

FIG. 7: 30-Day Rolling Average of Temperature and Humidity in Miami

Unlike Los Angeles, **Miami's temperature and humidity appear to move together**, both showing peaks roughly around the middle of the year, coinciding with hot and humid summer months typical of tropical climates. This means there is a direct dependence between temperature and humidity in Miami. (see 7)

In **Jerusalem** too, there is inverse relation between temperature and humidity. (see 8)

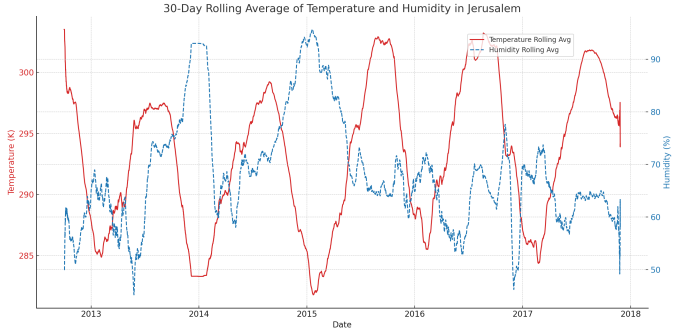

FIG. 8: 30-Day Rolling Average of Temperature and Humidity in Jerusalem

Now that we analyzed this, what are the factors that affect this trend and what can we conclude from this.
We see an inverse trend in Los Angeles and Jerusalem but a direct in Miami. This is due to its **Geographical Location and factors** around their locations. **Los Angeles and Jerusalem** are characterized by a **Mediterranean climate**, which typically exhibit warm to hot, dry summers and mild, wetter winters. During the summer, as temperatures rise, the relative humidity tends to decrease because warm air can hold more moisture than it generally does. This results in lower relative humidity levels during the hottest part of the day.
On the other hand when we look at Miami, it has a **tropical monsoon climate**, with hot and humid summers. The high temperatures facilitate the evaporation of water, which increases the moisture content in the air, thereby increasing the humidity. This explain the interdependence.

The reason for the sudden spikes is due to the **El Niño** effect, which for our dataset, occurred during 2015-2016.

El Niño caused **increased rainfall in Los Angeles**, **warmer winters in Miami**, and **wetter conditions** in Jerusalem. This is the reason why we see the spikes in those years in the graphs.

III. TIME SERIES ANALYSIS

In this section, we conduct our main time series analysis of **temperature and humidity data for Los Angeles**. We perform all the methods that we studied throughout this course to look at the underlying trends, seasonality analysis and end by trying to fit a model and forecast further values. We first look at the trend.

A. Long Term Trend

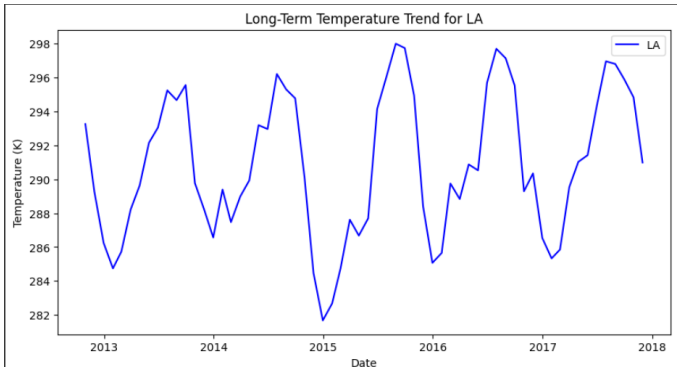

FIG. 9: Long-Term Temperature Trend for LA

The long-term temperature trend for Los Angeles indicates a slight upward trajectory over the years, suggesting a gradual increase in average temperatures. This is because of the global warming factors such as urban development and the Urban Heat Island effect, which tend to raise temperatures in metropolitan areas. ((see 9)

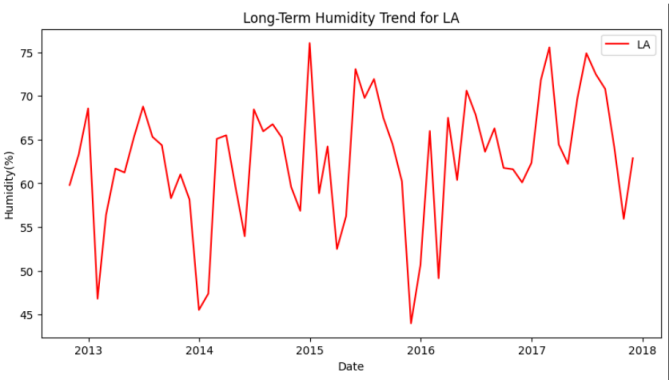

FIG. 10: Long-Term Humidity Trend for LA

The long-term humidity trend for Los Angeles (Figure 10) is relatively stable with minor variations over the years. This stability in humidity, despite rising temperatures, indicates the unique climatic conditions of Los Angeles, influenced by its coastal proximity which tends to moderate humidity levels. We can conclude that LA has a Coastal Climate by looking at this graph. (see 10)

B. Trend

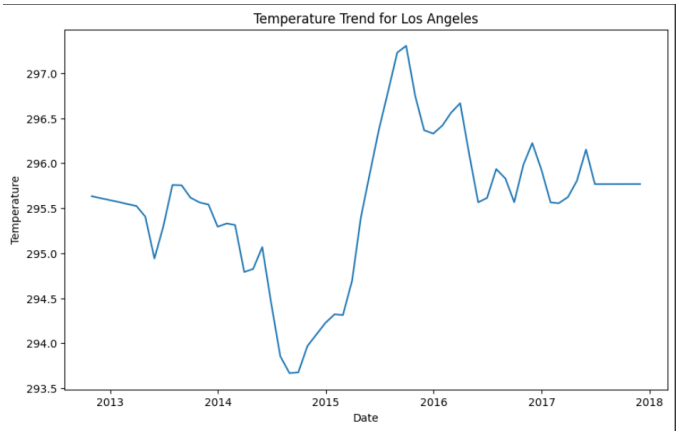

FIG. 11: Temperature Trend - LA

The trend analysis for temperature tells us that while there are fluctuations throughout the years, the general pattern shows **a subtle rise**. This indicates a broader climatic shift or local environmental changes impacting Los Angeles. The drop around 2015 is due to the **El Niño** effect that was discussed earlier. (see 11)

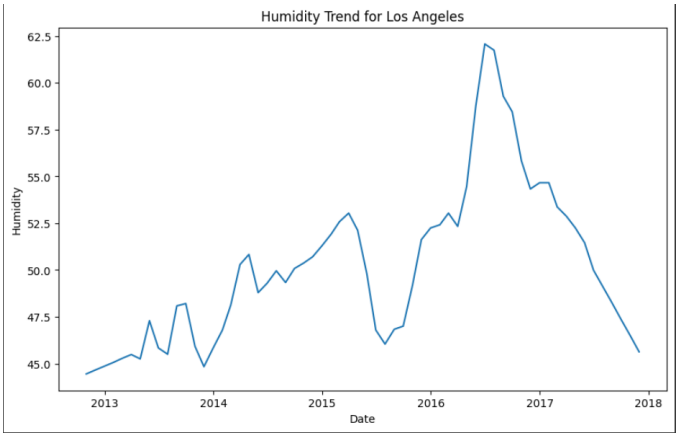

FIG. 12: Humidity Trend - LA

The detailed trend analysis for humidity (Figure 12) does not show significant long-term changes but does highlight seasonal variations, which remain consistent over the years, confirming the city's stable humid climatic conditions despite global climate variability. (see 12)

C. Seasonality

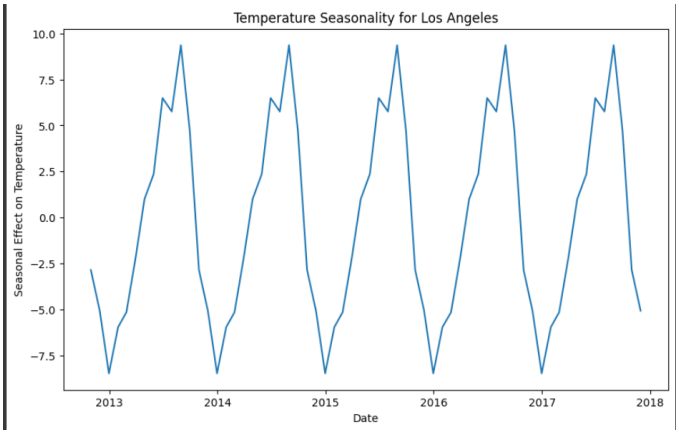

FIG. 13: Temperature Seasonality

This seasonality plot for temperature shows higher temperatures during the summer months and cooler temperatures during winter, following the **typical Mediterranean climate pattern** of Los Angeles. (see 13)

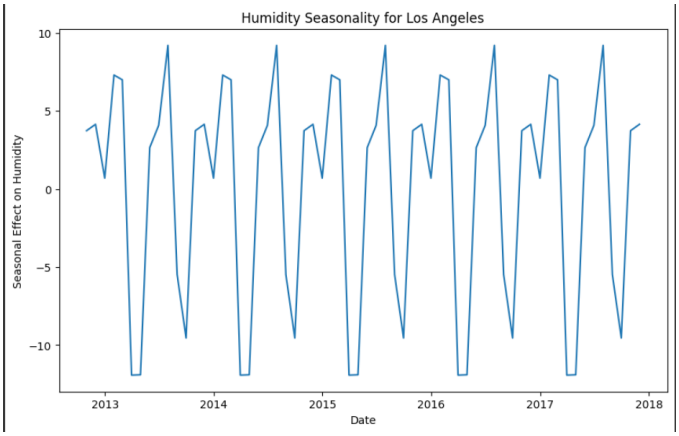

FIG. 14: Humidity Seasonality

The humidity seasonality in Los Angeles exhibits higher humidity during the cooler months and lower humidity during the summer. This **inverse relationship with temperature** that we saw earlier - cooler air can hold less moisture, leading to higher relative humidity when temperatures drop. (see 14)

IV. TIME SERIES DECOMPOSITION

A. Detrend and Deseasonal

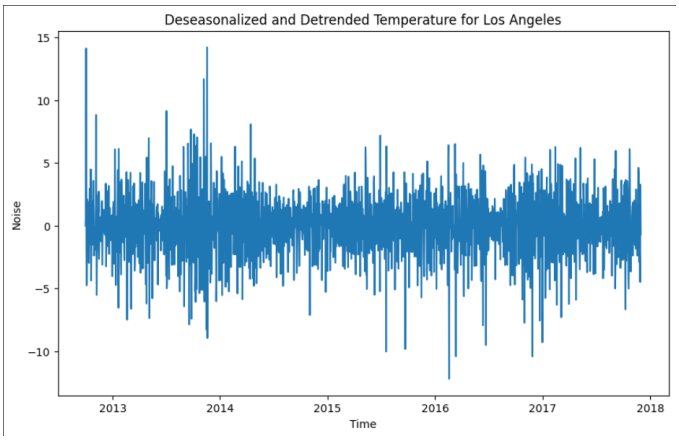

FIG. 15: Detrended and Deseasonalized Temperature

The detrended and deseasonalized temperature data illustrates residual fluctuations around a zero mean. This representation helps in identifying the irregularities and checking for stationarity. (see 15)

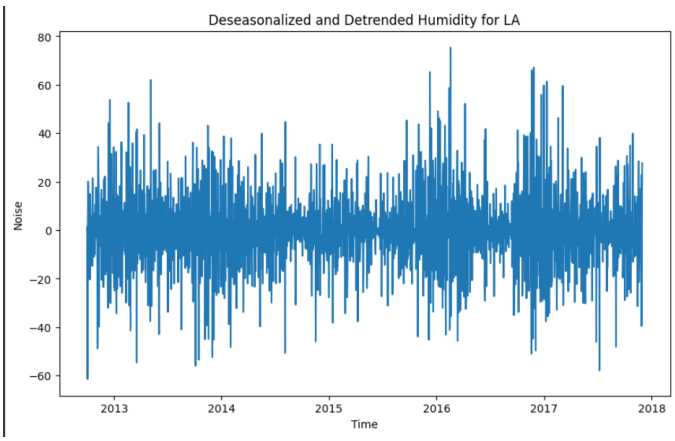

FIG. 16: Detrended and Deseasonalized Humidity

Figure (see 16) displays the detrended and deseasonalized humidity data, which oscillates around a baseline humidity value. This simplifies further analysis aimed at understanding residual anomalies and stationarity checks.

B. Rolling Mean and Standard Deviation Graph

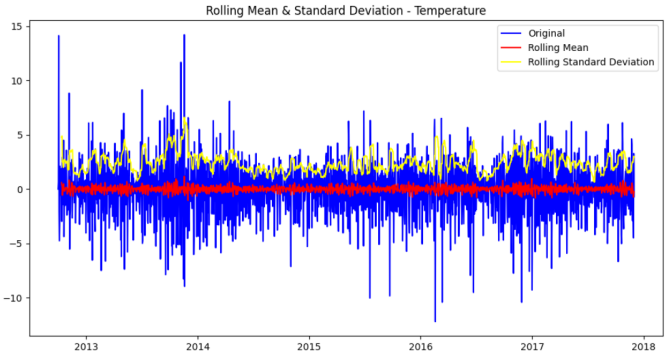

FIG. 17: Rolling Mean and Standard Deviation - Temperature

The rolling mean and standard deviation for temperature, as depicted in Figure (see 17), remain consistent and stable. This indicates that the temperature data, post-transformation, meets the criteria for stationarity, i.e. constant values

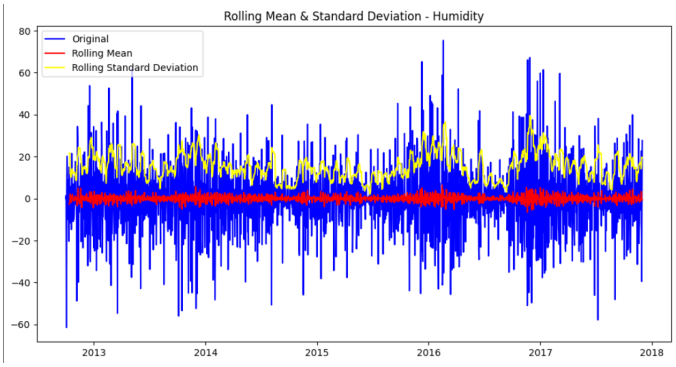

FIG. 18: Rolling Mean and Standard Deviation - Humidity

As shown in Figure (see 18), the rolling mean and standard deviation of humidity data exhibit minimal variation over time, which also indicates stationarity.

C. Autocorrelation Functions

The Autocorrelation function displayed in Figure 19 tell us about the temperature anomalies that can be modeled effectively using a **Moving Average process of order three** (MA(3)). The significant spikes in the Autocorrelation at lag 0, lag 2, and lag 3, with Autocorrelations returning towards zero at other lags, underscore the presence of short-term noise factors that impact the series only at these intervals.

Given the observed autocorrelation pattern, the temperature anomalies can be modeled as an MA(3) process,

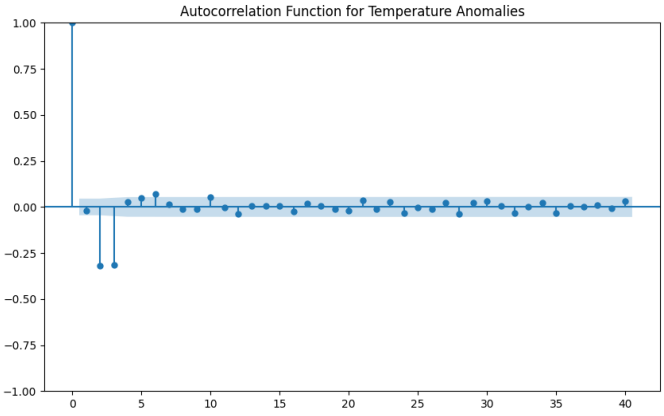

FIG. 19: ACF - Temperature

which is appropriately described by the equation:

$$X_t = \mu + \epsilon_t + \theta_2\epsilon_{t-2} + \theta_3\epsilon_{t-3} \tag{1}$$

where:

- μ is the mean of the series,
- ϵ_t are the white noise error terms,
- θ_2 and θ_3 are the parameters of the model corresponding to the significant autocorrelations at lags 2 and 3 respectively.

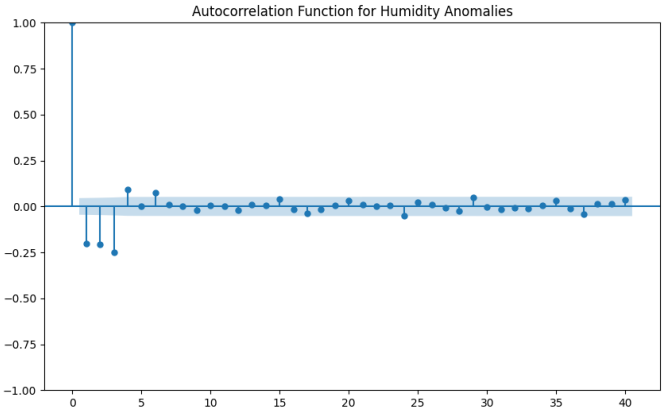

FIG. 20: ACF - Humidity

The ACF displayed in Figure (see 20) for humidity anomalies reveals a pattern indicative of a Moving Average process up to the third order (MA(3)). The significant autocorrelation at the initial lags (0, 1, 2, and 3) highlights the impact of error terms from the past three periods on the current value.

Given the autocorrelation pattern observed, the humidity anomalies can be effectively modeled as an MA(3) process. This is described by the equation:

$$X_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \theta_3\epsilon_{t-3} \tag{2}$$

where:

- μ is the mean of the series,
- ϵ_t are the white noise error terms,
- $\theta_1, \theta_2, \theta_3$ are the parameters of the model corresponding to the significant Autocorrelations at lags 1, 2, and 3 respectively.

V. MODEL

What is the main idea behind a **model** and finding the **best model** ? Our motive is to identify the most suitable statistical model that accurately captures the underlying patterns of the data. By selecting this optimal model, we aim to achieve accurate forecasts that will help us in making informed decisions based on predicted future values. This is important in weather forecasting as the day-to-day activities of a person depends on the environmental conditions. We plan to stay indoors when the forecast says "Heavy Rain for Tomorrow" or plan a nice excursion out when it says "Sunny Skies". Model fitting is important and here we have discussed how do we find the best model for our case.

A. Find the model

To identify the most suitable models for forecasting temperature and humidity anomalies, we employed the `pmdarima` Python library, which gives usthe process of ARIMA model selection. This library implements a stepwise search across defined model parameters, which gives us the best fitting configuration for our SARIMAX model.

For each dataset, `auto_arima` function was configured to evaluate both non-seasonal and seasonal components, given the daily frequency of data collection. The function iteratively fits numerous ARIMA models by varying the order of differencing (d), the number of autoregressive terms (p), and the number of moving average terms (q), including possible seasonal terms.

We have considered various differencing orders to make our time series stationary; in that brute force approach, we get **first-order differencing as best** because it has nearly zero mean with very low tolerance, which means it has very close to zero mean, and in our stationarity analysis it is best to get the zero mean processes.

We also confirmed this using the SARIMAX grid search's predict best-fit model, that first-order differencing can remove the seasonal part of the time series.

B. The SARIMAX model

The Seasonal AutoRegressive Integrated Moving Average with eXogenous variables model (SARIMAX) extends the ARIMA model by incorporating both seasonal

differences and exogenous factors. This model is particularly useful in scenarios where data exhibit seasonal patterns and can be influenced by external variables.

Mathematical Representation

The SARIMAX model equation is given by:

$$\begin{aligned} Y_t = \alpha + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \\ + \sum_{i=1}^{P} \Phi_i Y_{t-si} + \sum_{i=1}^{Q} \Theta_i \epsilon_{t-si} \\ + \beta X_t + \epsilon_t \end{aligned} \tag{3}$$

where:

- Y_t is the observed time series at time t.
- α is a constant term.
- ϕ_i are the coefficients for the non-seasonal autoregressive (AR) terms.
- θ_i are the coefficients for the non-seasonal moving average (MA) terms.
- Φ_i are the coefficients for the seasonal AR terms.
- Θ_i are the coefficients for the seasonal MA terms.
- si is the seasonal period multiplied by the lag i.
- βX_t represents the influence of exogenous regressors.
- ϵ_t is the error term at time t.

C. Fitting the model

The SARIMAX parameters were selected based on a combination of statistical tests, including ADF tests for stationarity and ACF and PACF plots for determining appropriate lags. For humidity, a SARIMAX(3,0,0)(2,1,0)[12] model was fitted, reflecting three autoregressive terms and a seasonal differencing of one at a yearly cycle. Temperature data was modeled similarly, and here are the results.

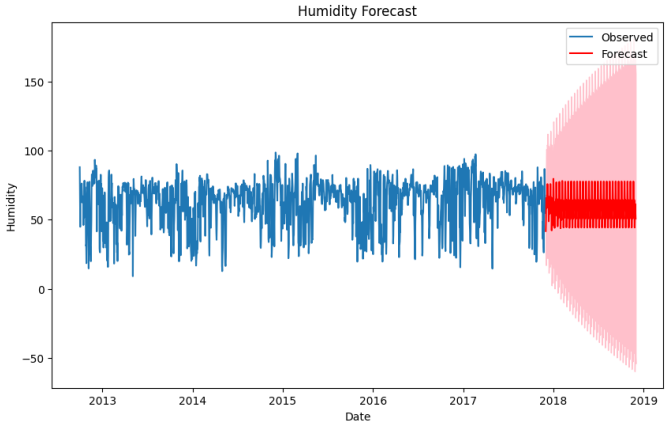

FIG. 21: Humidity Forecast

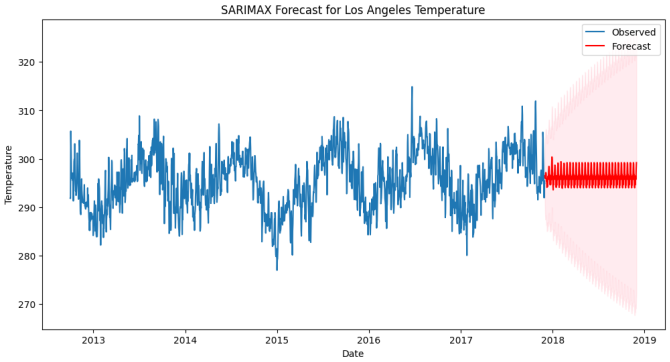

FIG. 22: Temperature Forecast

The above are the Temperature (Fig. 22) and Humidity (Fig. 21) Forecasts. The forecasting of our time series is not as expected because the internal library of Python uses grid search to get the model that suits our time series best; as grid search has limited parametric space, it may not be able to go to its full potential for the best-fit model for our time series.

We have complex relational time series that may require better hyperparameter tuning to get the best-fit model. The grid search may lack that approach, but we can predict for a shorter run that may fulfill our requirements.

VI. CONCLUSIONS

This project successfully applied SARIMAX models to analyze and forecast temperature and humidity trends in Los Angeles from the data that we had of 2012-17. The study revealed distinct patterns in temperature and humidity related to seasonal variations and urban density. We saw that the city of Los Angeles exhibited higher humidity and temperature levels, largely influenced by human activities such as industrial outputs and vehicu-

lar emissions, which would be less when compared to a non-urbane atmosphere.

The use of SARIMAX models enabled a detailed understanding of these temperature and humidity trends by accounting for the trends and seasonal changes, and impact of exogenous variables such as urban development and green space distribution.

However, the study also encountered limitations. The models sometimes failed to predict abrupt changes in humidity brought on by sporadic events like sudden weather shifts or unseasonal rainfalls . This highlighted a need for refining the models to better accommodate irregular climatic phenomena.

In conclusion, employing ARIMA and SARIMAX models has provided valuable insights into the dynamics of temperature and humidity. These models proved effective in capturing long-term and seasonal trends, yet they require enhancements to improve the accuracy of predictions against atypical events. Future research could explore integrating more dynamic and complex models or incorporating broader climatic data sets to refine predictions.

Final remarks , we learned the importance of these metrics in our daily lives and how they can be affected by even the smallest of the factors can have a huge impact on the weather conditions which has a huge impact on the individual, city, nation or world level. This is our contribution to the community and making it an "easier" place to live and work in!

VII. PRESENTATION VIDEO

Here is the link to our Presentation Video : Video Presentation

[1] D. Beniaguev, *Historical hourly weather data 2012-2017*, Kaggle (2017), available online: https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data.

[1]