INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

ROBERT GORDON UNIVERSITY ABERDEEN

Artificial Intelligence and Data Science

Module Leader: Mohamed Ayoob

CM 2604 – Data Engineering

Assignment type: Individual Coursework

Student Name: M.Hasinindu Jayashan De Silva

IIT ID – 20211295

RGU ID – 2312548

Contents

Table of Figures

1. Introduction

Formaldehyde (HCHO) is a volatile organic compound widely recognized for its implications on environmental pollution and public health. Monitoring its levels across various geographical locations is crucial for assessing air quality and implementing effective environmental protection policies. This project focuses on a detailed analysis of tropospheric HCHO concentrations across different cities in Sri Lanka.

Dataset Overview:

The dataset used in this study comprises measurements of HCHO levels collected from seven distinct urban and rural locations within Sri Lanka over a period spanning from January 2019 to December 2023. These locations include Colombo Proper, Deniyaya Matara, Nuwara Eliya Proper, Bibile Monaragala, Kurunegala Proper, Jaffna Proper, and Kandy Proper. The data, derived from satellite observations combined with ground-based monitoring systems, includes over 12,782 recorded observations, each quantifying the HCHO concentration, location, and the exact date of measurement.

2. Data Preprocessing

2.1 Initialize Spark Session

```
from pyspark.sql import SparkSession
spark = SparkSession.builder \
    .appName("HCHO Data Analysis") \
    .getOrCreate()

spark
```

SparkSession - in-memory

SparkContext

Spark UI

Version
v3.5.1
Master
local[*]
AppName
HCHO Data Analysis

2.2 Load the Data

Load 'col_mat_nuw_output.csv' file

```
col_mat_nuw_df = spark.read.csv('Dataset/col_mat_nuw_output.csv',
header=None, inferSchema=True)
col_mat_nuw_df.show()
```

Load 'mon_kur_jaf_output.csv' file

```
mon_kur_jaf_df = spark.read.csv('Dataset/mon_kur_jaf_output.csv',
header=None, inferSchema=True)
mon_kur_jaf_df.show()
```

Load 'kan_output.csv' file

```
kan_df = spark.read.csv('Dataset/kan_output.csv', header=None,
inferSchema=True)
```

2.2.1 Changing Header Names

```
column_names = ['HCHO', 'Location', 'Current date', 'Next date']
```

```
for i, new_name in enumerate(column_names):
    col_mat_nuw_df =
col_mat_nuw_df.withColumnRenamed(col_mat_nuw_df.columns[i], new_name)

col_mat_nuw_df.show()
```

```
+--------------------+--------------+------------+----------+
|                HCHO|      Location|Current date| Next date|
+--------------------+--------------+------------+----------+
|1.969834395781014...|Colombo Proper|  2019-01-01|2019-01-02|
|2.625522171968594...|Colombo Proper|  2019-01-02|2019-01-03|
|9.852118897938794E-5|Colombo Proper|  2019-01-03|2019-01-04|
|2.099320518114242E-4|Colombo Proper|  2019-01-04|2019-01-05|
```

The above code shows the way that I have changed the header names of the 'col_mat_nuw_output.csv' file. Furthermore, I have used a similar code to change the header names of the other 2 csv files.

2.2.2 Merging Datasets

The below set of codes show how I have merged the 3 datasets and exploration of the descriptive statics.

```
combined_df = col_mat_nuw_df.union(mon_kur_jaf_df).union(kan_df)
```

```
combined_df.printSchema()
```

```
root
 |-- HCHO: double (nullable = true)
 |-- Location: string (nullable = true)
 |-- Current date: date (nullable = true)
 |-- Next date: date (nullable = true)
```

```
num_rows = combined_df.count()
num_cols = len(combined_df.columns)

print(f"Number of rows: {num_rows}")
print(f"Number of columns: {num_cols}")
```

```
Number of rows: 12782
Number of columns: 4
```

```
combined_df.describe().show()
```

```
+-------+--------------------+--------------------+
|summary|                HCHO|            Location|
+-------+--------------------+--------------------+
|  count|                7918|               12782|
|   mean|1.192778916513748...|                NULL|
| stddev|9.322341209771851E-5|                NULL|
|    min|-3.52473024357239...| Bibile, Monaragala|
|    max|8.997101837438971E-4|Nuwara Eliya Proper|
+-------+--------------------+--------------------+
```

After the exploration of the descriptive statics the merged dataset was saved as a csv file

```
pandas_df = combined_df.toPandas()

output = 'Dataset/combined_df.csv'

pandas_df.to_csv(output, index=False)

print(f"Pandas dataframe saved to {output}")
```

2.3 Cleaning the Data

Distribution of the HCHO Reading of the combined dataset

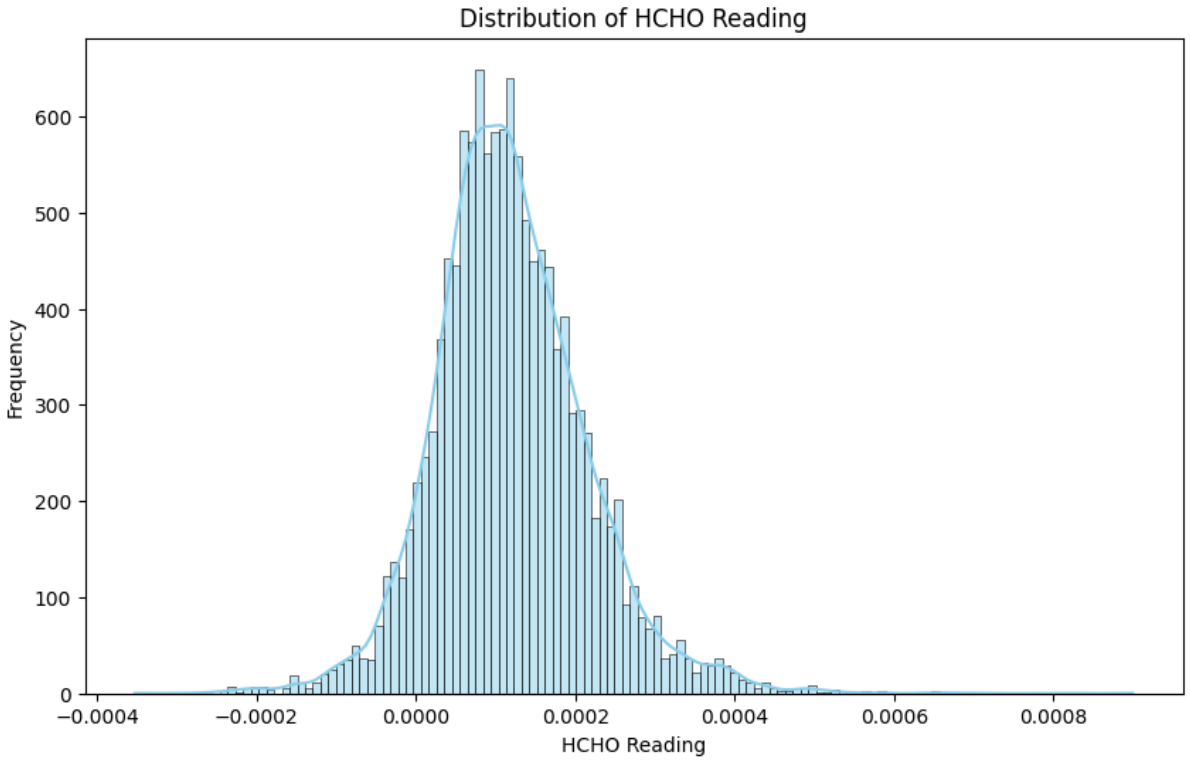


Figure 1: Distribution of HCHO Reading

2.3.1 Checking for Duplicates

```
total_rows = combined_df.count()
distinct_rows = combined_df.distinct().count()

print(f"Total rows: {total_rows}")
print(f"Distinct rows: {distinct_rows}")
if total_rows > distinct_rows:
    print("There are duplicates in the DataFrame.")
else:
    print("No duplicates found.")
```

```
Total rows: 12782
Distinct rows: 12782
No duplicates found.
```

2.3.2 Checking Null Values

```
from pyspark.sql.functions import col, count, when, isnull

null_counts = combined_df.select([count(when(isnull(c), c)).alias(c) for c
in combined_df.columns])
null_counts.show()
```

```
+----+--------+------------+---------+
|HCHO|Location|Current date|Next date|
+----+--------+------------+---------+
|4864|       0|           0|        0|
+----+--------+------------+---------+
```

As per the output there are 4864 total null values in the whole dataset. I used forward filling and backward filling to handle the null values. The below chart shows the null value count after handling them.

```
+----+--------+------------+---------+
|HCHO|Location|Current date|Next date|
+----+--------+------------+---------+
|   0|       0|           0|        0|
+----+--------+------------+---------+
```

2.3.3 Handling Outliers

Boxplot of the combined dataset before handling outliers.

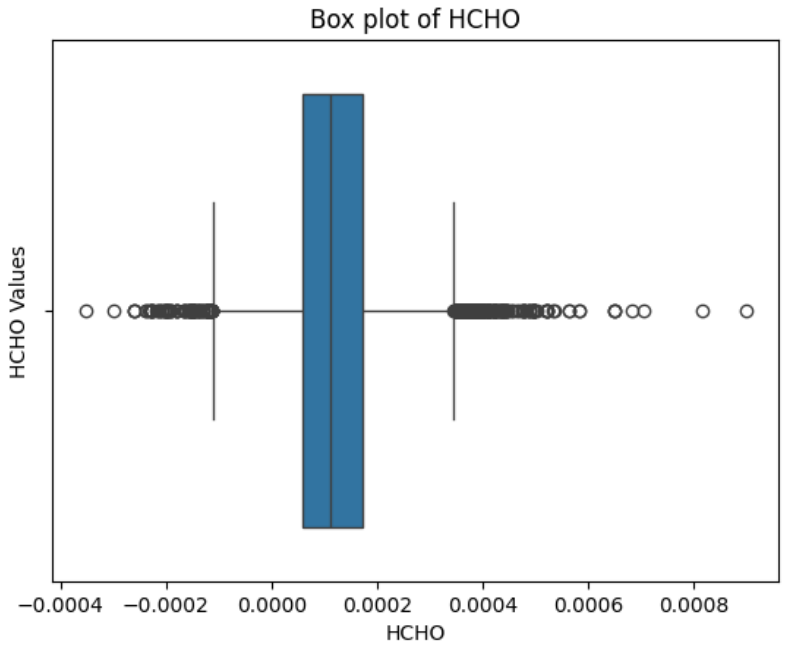


Figure 2: Boxplot of HCHO before handling outliers

Boxplot of the combined dataset after handling outliers.

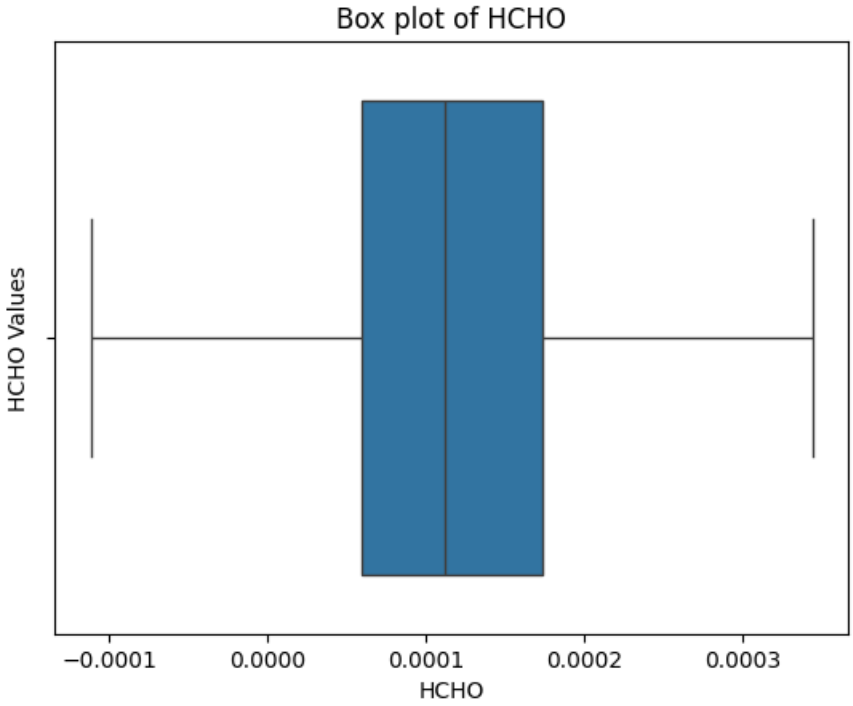


Figure 3: Boxplot of HCHO after handling outliers

2.3.4 Checking Unique Values

```
for unique_value in sorted(pandas_df['Location'].unique()):
    print(unique_value)
```

```
Bibile, Monaragala
Colombo Proper
Deniyaya, Matara
Jaffna Proper
Kandy Proper
Kurunegala Proper
Nuwara Eliya Proper
```

This is the output for the Location column. I have checked for the unique values in each column.

3. Spatio-Temporal Analysis

3.1 Seasonal variations

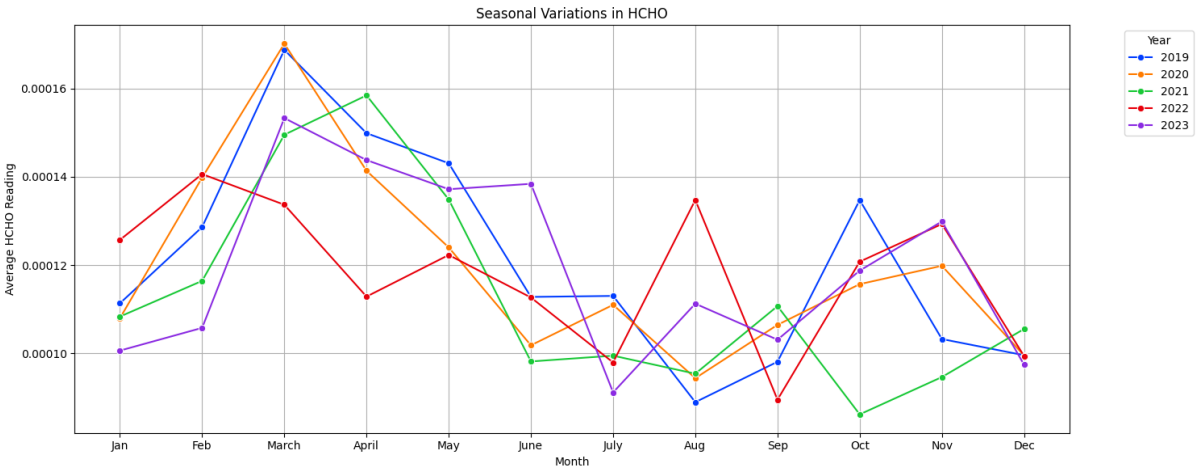


Figure 4: Seasonal Variation in HCHO Levels

This graph shows the seasonal variations in formaldehyde (HCHO) concentrations over several years, from 2019 to 2023. Each line in the graph represents a different year. The horizontal axis shows the months of the year, from January to December, indicating that the data is observed and compared monthly. The vertical axis represents the average readings of HCHO concentrations.

Seasonal patterns:

There appears to be a seasonal pattern, with peaks and troughs occurring around the same time each year, which suggests a possible correlation between the time of year and HCHO levels. Certain months, like April to June (spring to early summer), generally show higher average HCHO readings across multiple years, indicating that these months might have specific factors contributing to higher formaldehyde concentrations. Months from August to October (late summer to autumn) usually exhibits a decrease in HCHO levels There are variations in the patterns from year to year, which could be due to several factors including changes in environmental policies, economic activities, or even climatic conditions affecting the presence of HCHO in the environment.

Month to month fluctuations:

From the graph, we can see that certain months, like April, exhibit high levels of HCHO across multiple years, showing that this month has specific environmental factors that contribute to higher emissions or reduced dispersion of HCHO. (New year season). Some months, such as

July and August, show less consistency across the years, indicating that other variable factors might be influencing the HCHO readings during these months.

Overall trends:

While the overall seasonal trend appears consistent with higher concentrations in certain months, the year 2022 stands out with generally higher HCHO levels throughout the year when compared to other years, indicating that there may have been an overall increase in HCHO concentrations in that particular year.

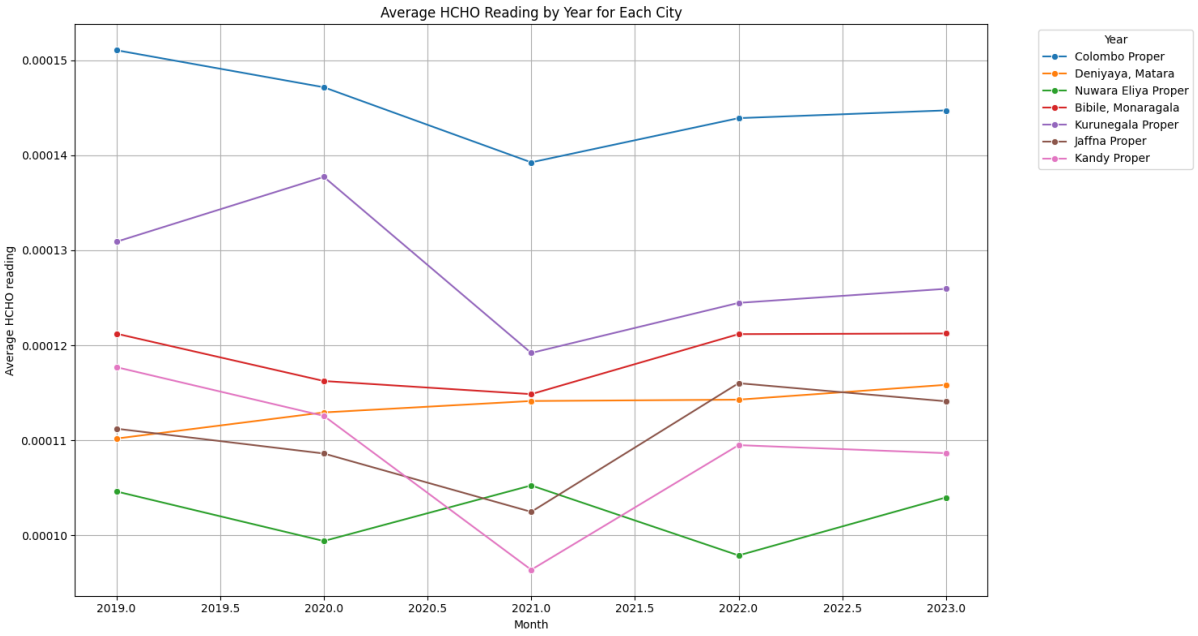


Figure 5: Average HCHO Reading by Year for Each City

This is a line graph that shows the seasonal variation in HCHO levels across 7 cities over several years. Each line represents a different city, as indicated by the color-coded legend. The lines show how the average HCHO readings change over time, suggesting a seasonal pattern in the data.

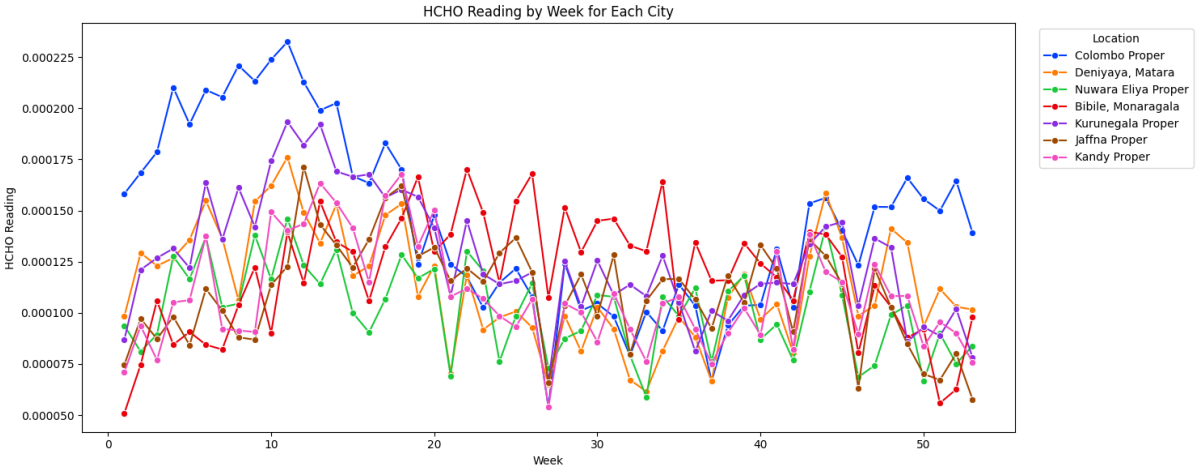


Figure 6: HCHO Reading by week for each city

The chart is a line graph titled "HCHO Reading by Week for Each City", displaying the weekly HCHO (formaldehyde) concentration readings across various cities. Each line represents the HCHO readings for a different city. There's significant variability in HCHO readings both within each city over time and between different cities. Some cities, such as Colombo Proper (blue line), show higher peak HCHO readings compared to other cities.

3.2 Long term trends

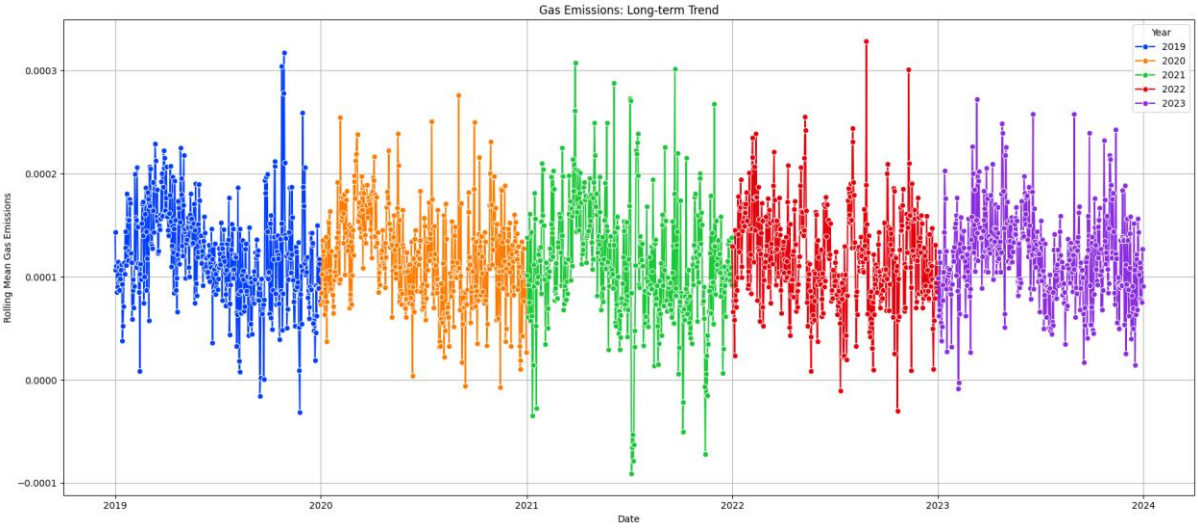


Figure 7: Gas Emission - Long Term Trends

The chart is a visual representation of gas emissions over time, specifically focusing on the long-term trend across multiple years (2019 to 2023). Data points are plotted for each interval and connected by lines, which seem to represent the rolling average of gas emissions at each point in time. The lines are color-coded by year, which allows for a comparison of emission levels year over year. By looking at the different colored lines, you can compare emission levels

across the years. For example, it seems that the variability and the range of the emissions are quite similar year over year, but there might be slight differences in the levels at comparable times of each year.

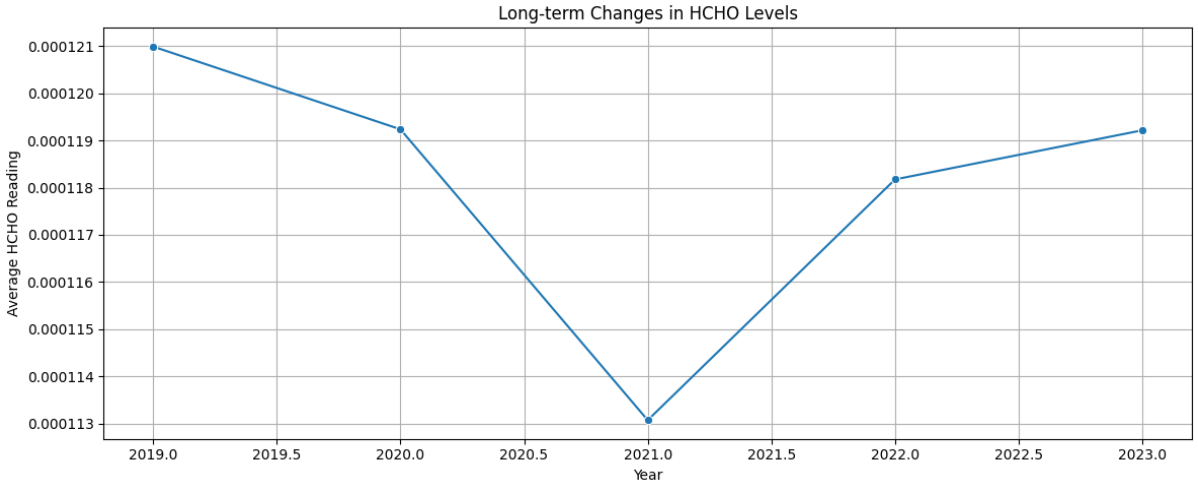


Figure 8: Long Term Trends in HCHO Levels

This chart shows the long-term changes in average HCHO (formaldehyde) levels over a period from the beginning of 2019 to the end of 2023. Starting in early 2019, there is a downward trend in HCHO levels until early 2021, where the lowest point is reached. From early 2021 to the end of 2023, there is an upward trend, indicating that average HCHO levels are increasing over this period. The V-shaped curve in the data might reflect a specific event's impact on HCHO levels, such as the COVID-19 pandemic, where initial strict measures could have reduced emissions followed by a recovery phase.

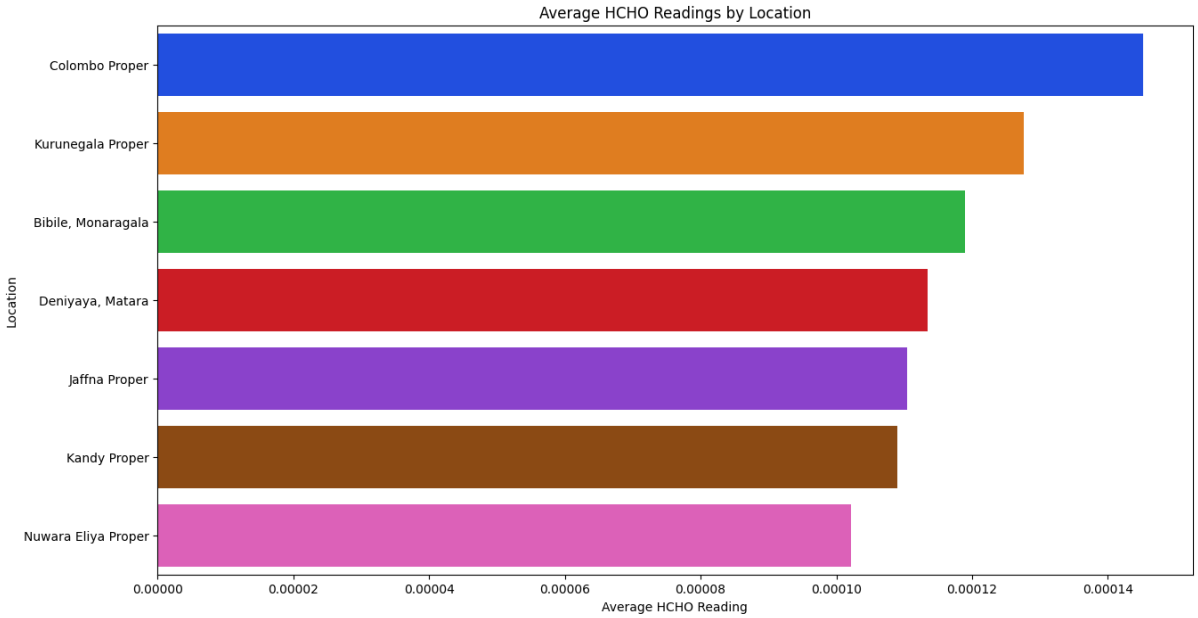


Figure 9: Average HCHO Reading by city

The chart is a horizontal bar graph showing average HCHO readings by location. Colombo Proper has the highest average HCHO reading among the listed locations, which might suggest more significant sources of HCHO emissions, such as industrial activities or traffic congestion. Nuwara Eliya Proper has the lowest average reading, which could be due to various factors like cleaner air, less industrial activity, or fewer emissions sources in general.

3.3 Trends across cities

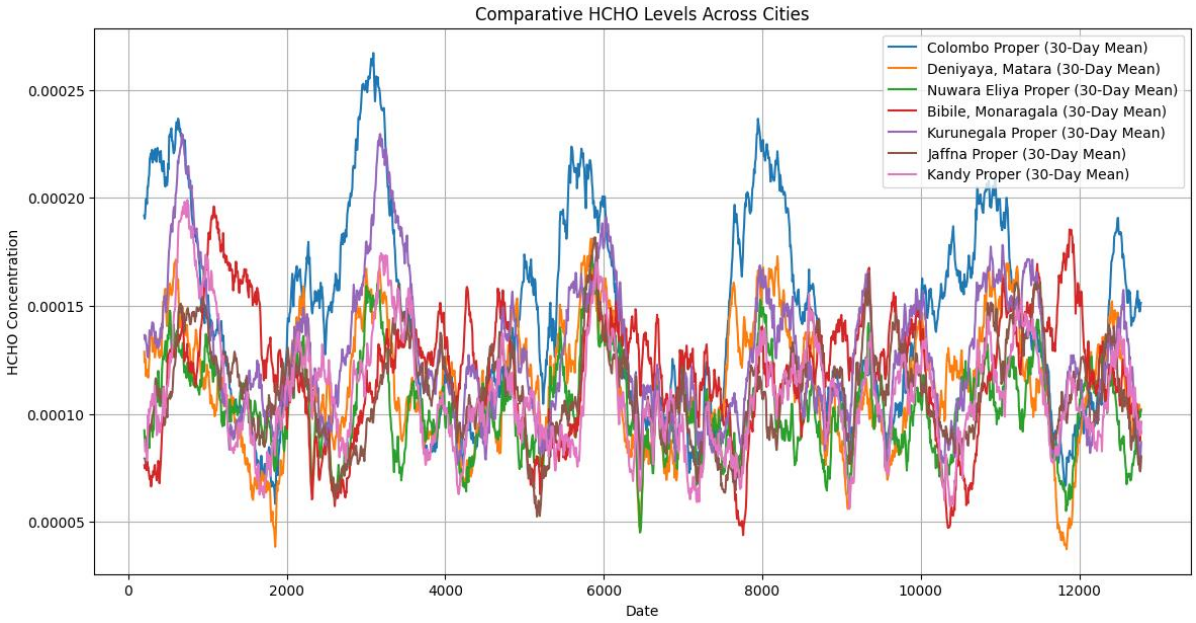


Figure 10: Comparative HCHO Levels Across Cities

This chart displays the comparative HCHO levels across various cities, with each cities data smoothed by a 30 day rolling mean to identify broader trends over the dates represented on the x-axis. Each line represents a city's 30-day rolling average of HCHO levels.

Tends and Patterns: By comparing the peaks and valleys of each line, it's possible to identify times when HCHO concentrations were universally higher or lower across most cities, which could indicate widespread environmental events or seasonal effects. For example, multiple cities show peaks around the same points on the x-axis, suggesting a seasonal or periodic event affecting HCHO levels simultaneously.

Bibile, Monaragala

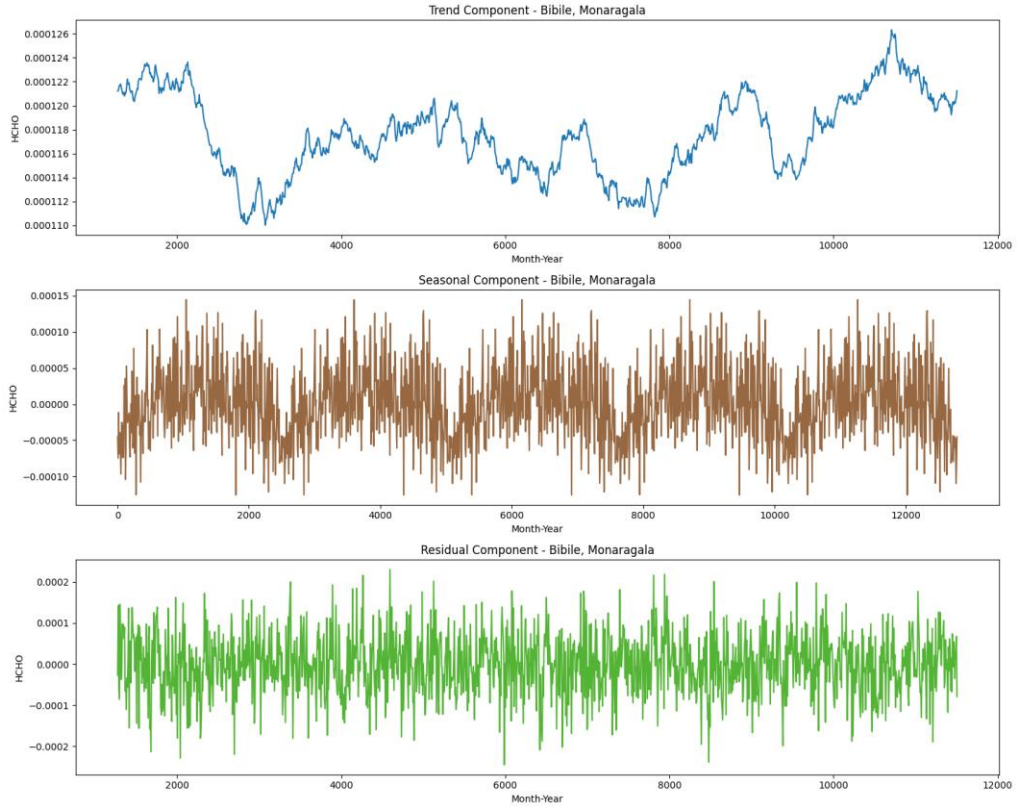


Figure 11: Bibile, Monaragala - Trend, Seasonal, Residual Component

Colombo proper

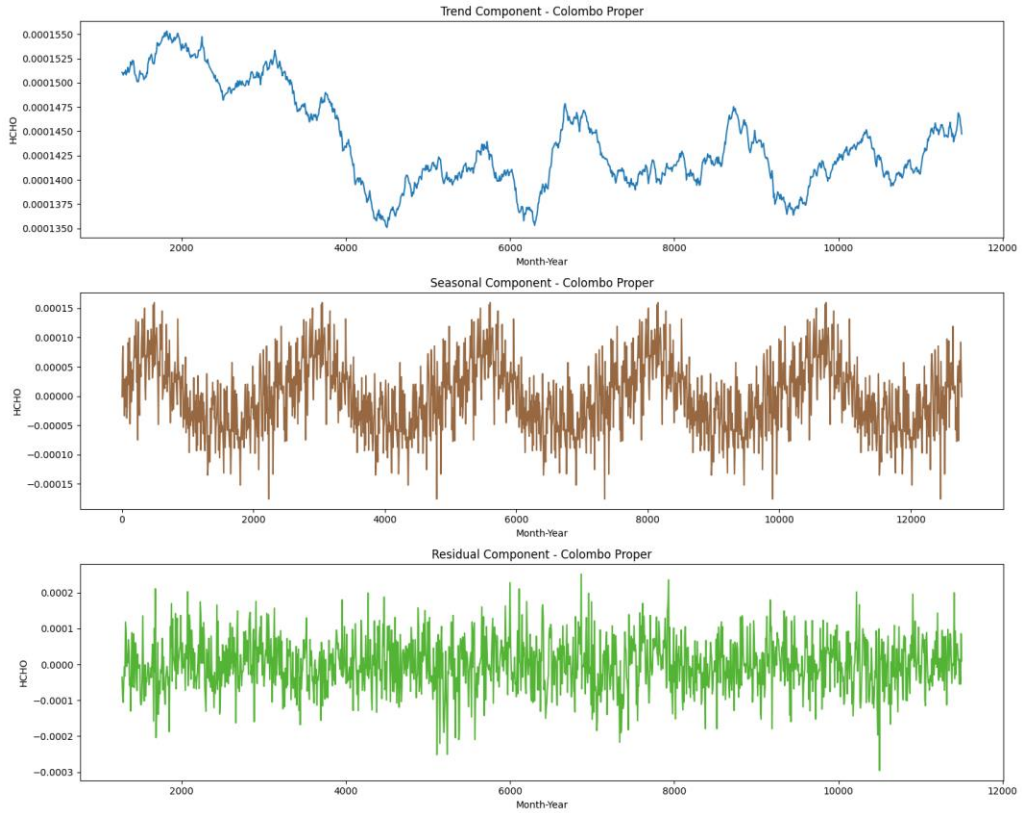


Figure 12: Colombo Proper - Trend, Seasonal, Residual Component

Deniyaya, Matara

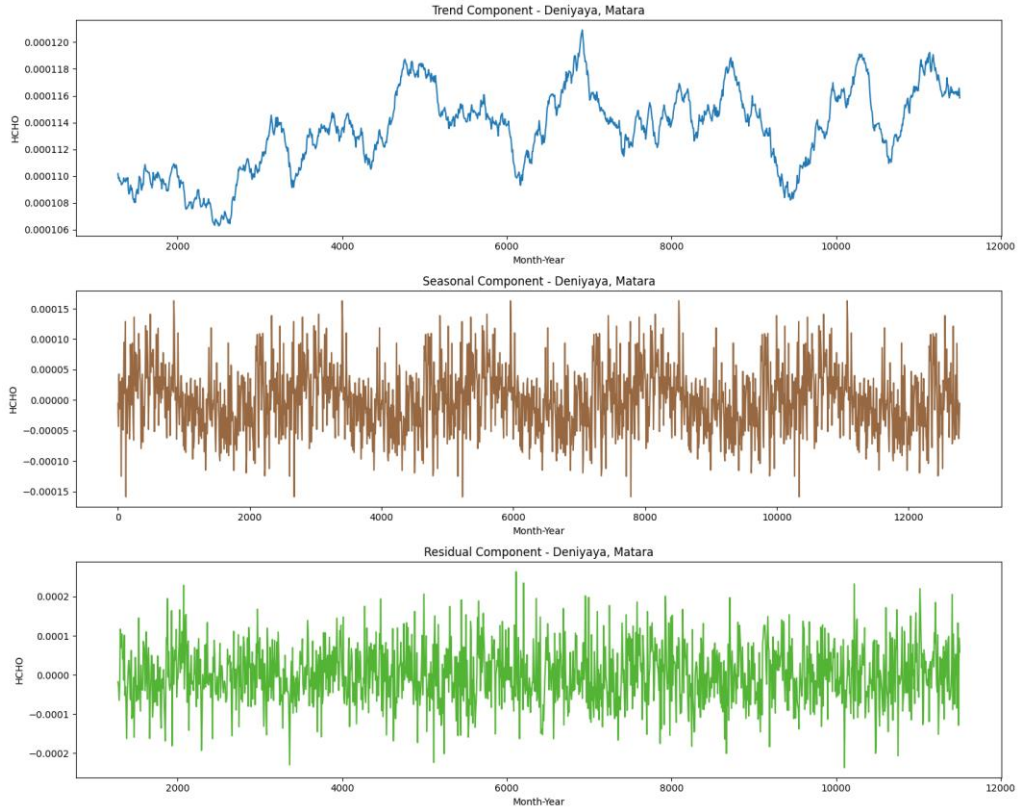


Figure 13: Deniyaya, Matara - Trend, Seasonal, Residual Component

Jaffna Proper

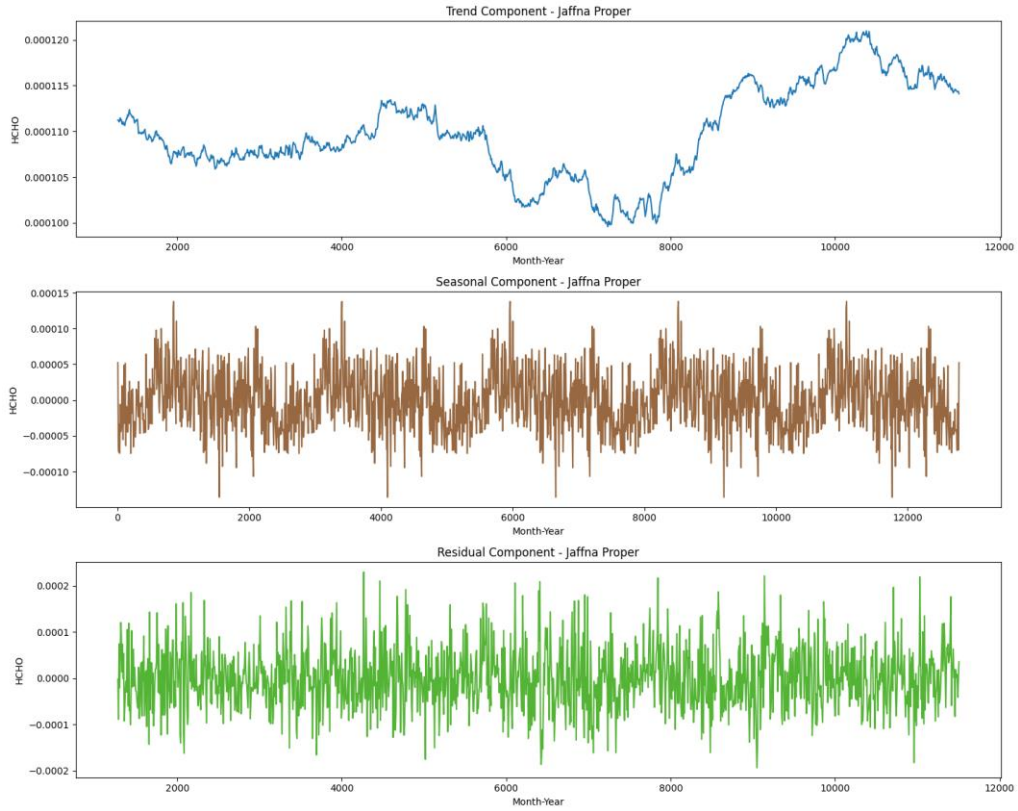


Figure 14: Jaffna Proper - Trend, Seasonal, Residual Component

Kandy Proper

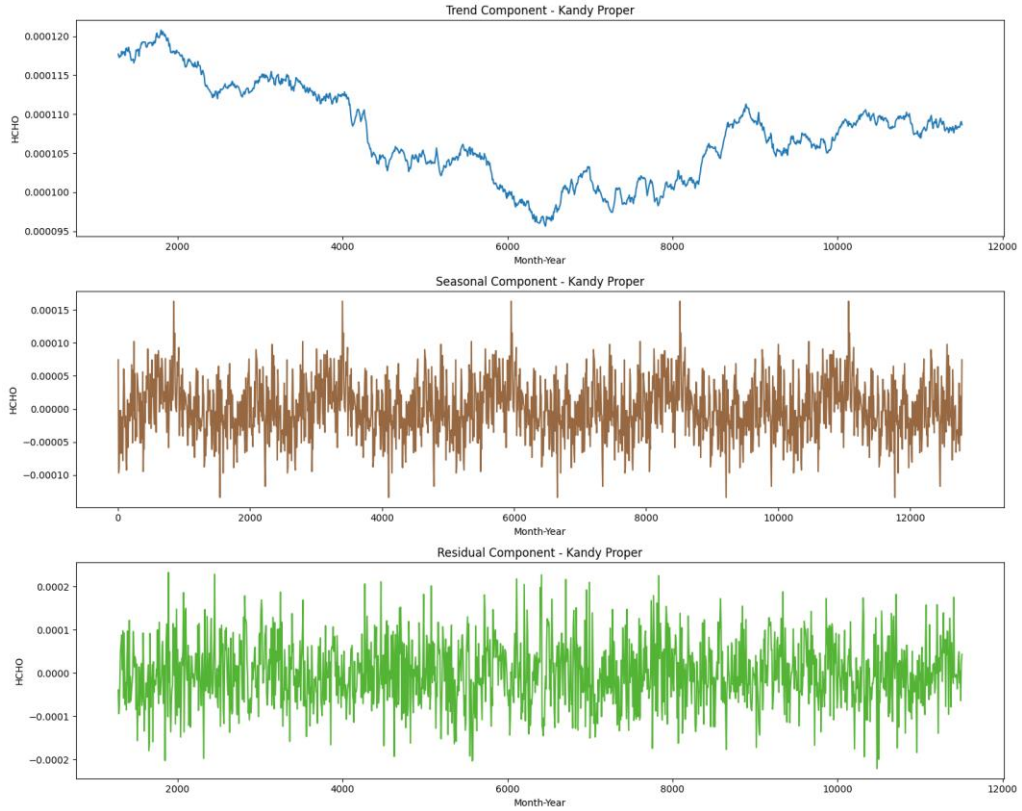


Figure 15: Kandy Proper - Trend, Seasonal, Residual Component

Kurunegala Proper

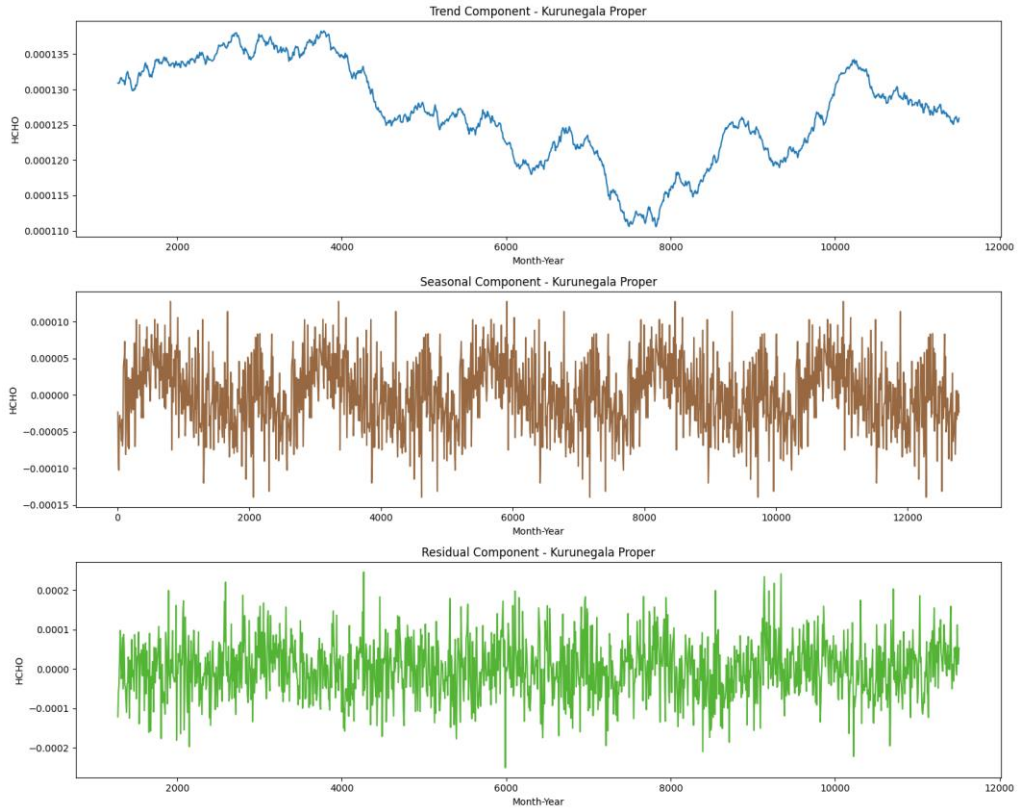


Figure 16: Kurunegala Proper - Trend, Seasonal, Residual Component

Nuwara Eliya Proper

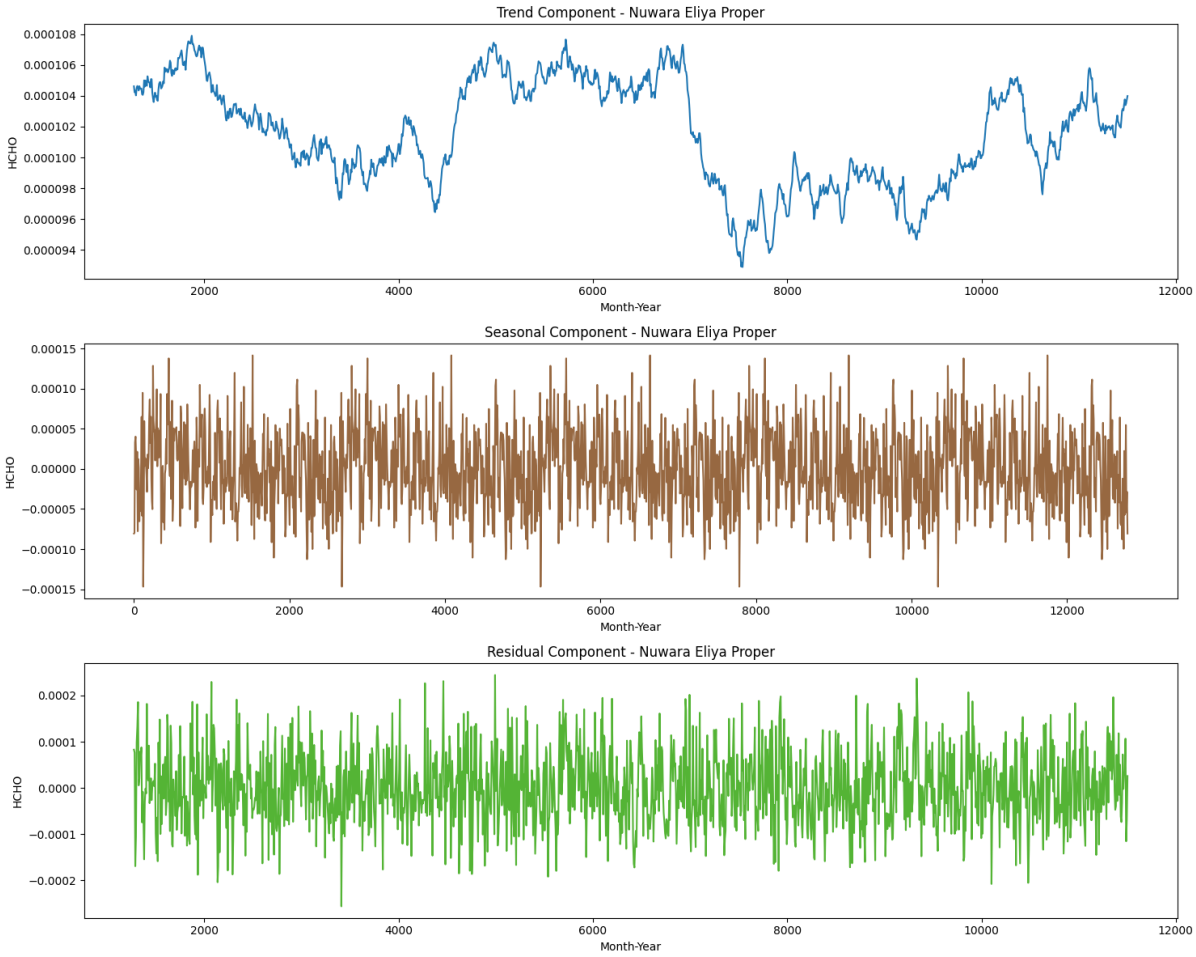


Figure 17: Nuwara Eliya Proper - Trend, Seasonal, Residual Component

The charts provided are time series decompositions of HCHO levels in various cities. Each city's data has been decomposed into three components: trend, seasonal, and residual. The trend component shows the long-term progression of HCHO levels, the seasonal component shows the repeating short-term cycle within the data, and the residual component shows the irregularities or noise that cannot be attributed to the trend or seasonality.

Trend Component:

- All cities show some variation in the trend over time, but the overall direction of the trend can vary from one city to another.
- In some cities, there is a noticeable upward or downward trend at certain periods, suggesting a change in HCHO levels over the long term.

Seasonal Component:

- The seasonal patterns appear to have a consistent fluctuation width, indicating a regular cycle of increases and decreases in HCHO levels within each year.

- However, the amplitude and pattern of the seasonal fluctuations can differ between cities, possibly due to the different climatic or environmental factors specific to each location.

Residual Component:

- The residual component represents random fluctuations that cannot be explained by the trend or seasonality. This could include random measurement errors, unexplained variance, or other stochastic processes affecting the HCHO levels.
- Some cities might exhibit larger residuals than others, which could indicate more random variability or noise in the HCHO level data.

Comparative Analysis:

- When comparing the trend components, you might notice that some cities have trends with more pronounced changes over time, suggesting significant long-term changes in HCHO levels.
- Comparing the seasonal components can reveal differences in the seasonal impact on HCHO levels between cities. Cities closer in geographic location may have similar seasonal patterns due to similar weather patterns.
- The residual components can be compared to evaluate the stability and predictability of the HCHO levels. A larger spread in the residuals might suggest that the city's HCHO levels are influenced by more erratic factors.

3.4 Covid-19 lockdown impact

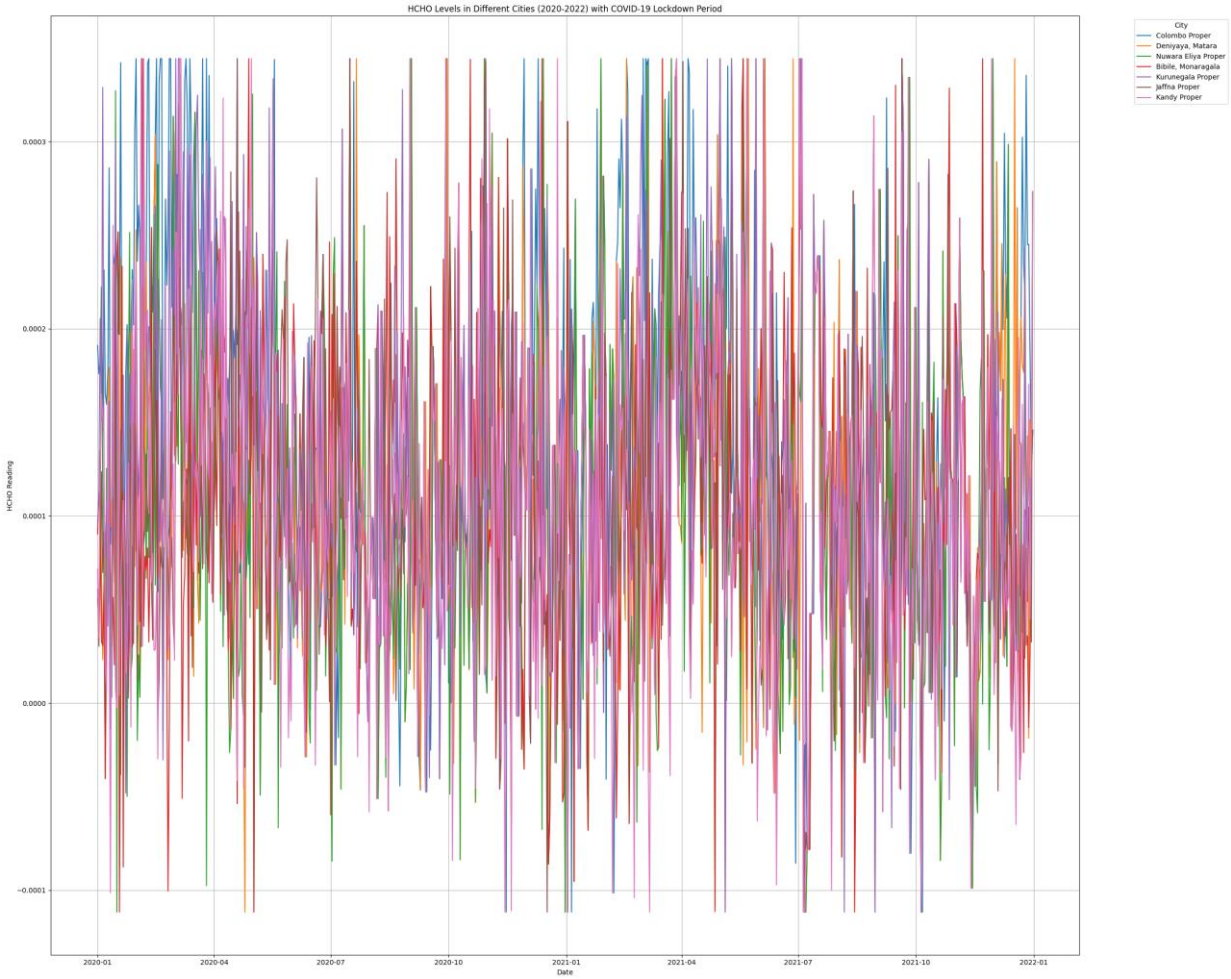
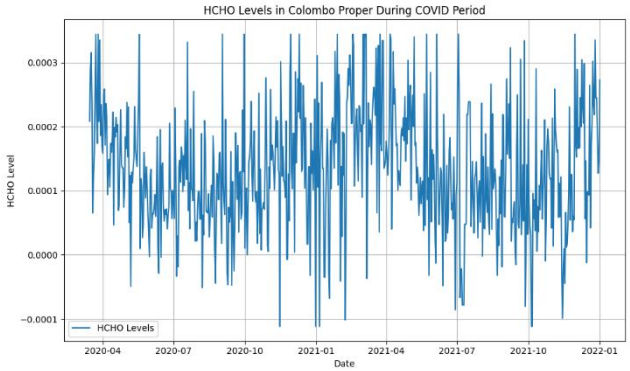
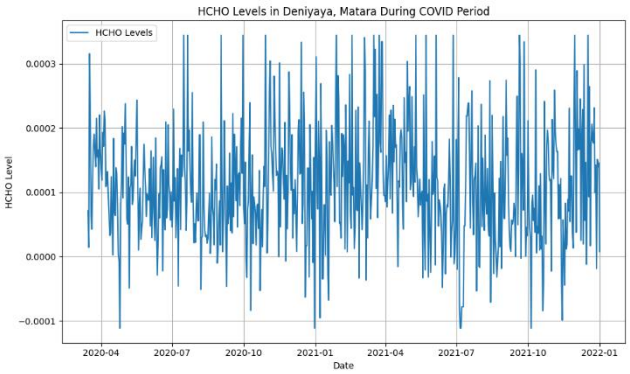


Figure 18: HCHO Levels in Different Cities (2020-2022) with COVID-19 Lockdown Period

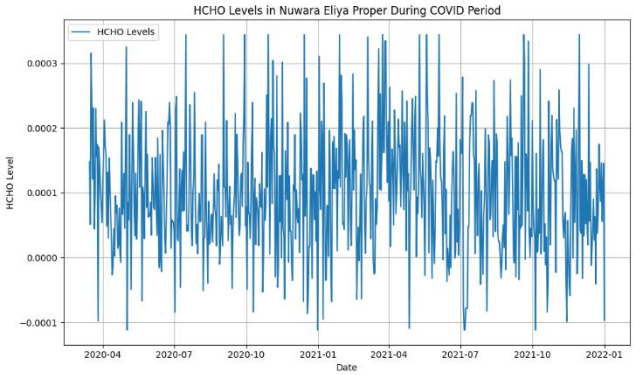
This graph represents the levels of HCHO in different cities over time, specifically from 2020 to 2022, which includes the Covid-19 lockdown period. The graph suggests that HCHO levels have varied considerably during the covered period, and differences between cities can be observed.



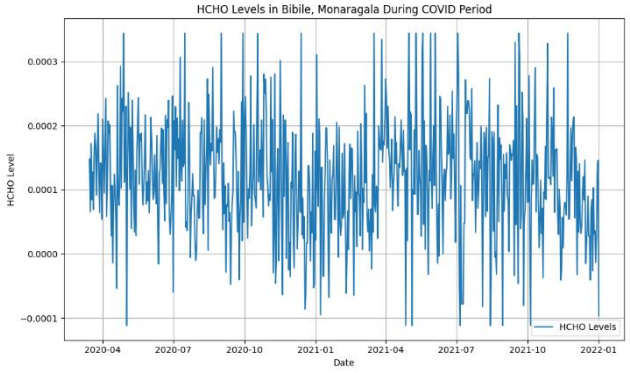
Mean HCHO levels in Colombo Proper:
Pre-COVID: 0.00016419085487377996
During COVID: 0.00013376558370895852
Post-COVID: 0.00014413998728524262



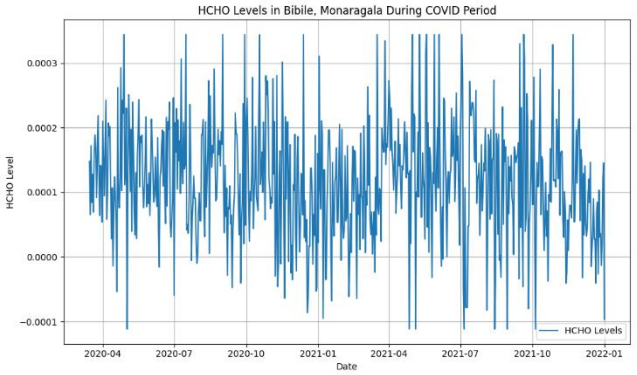
Mean HCHO levels in Deniyaya, Matara:
Pre-COVID: 0.00011438064960685291
During COVID: 0.00011095104869206006
Post-COVID: 0.00011520643294784729



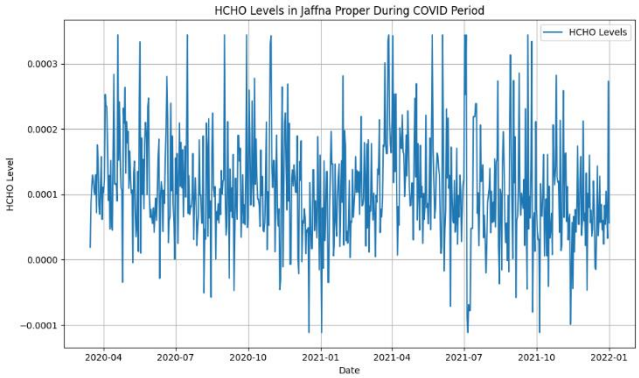
Mean HCHO levels in Nuwara Eliya Proper:
Pre-COVID: 0.00010688492720328902
During COVID: 0.0001002455795582544
Post-COVID: 0.00010120517334799341



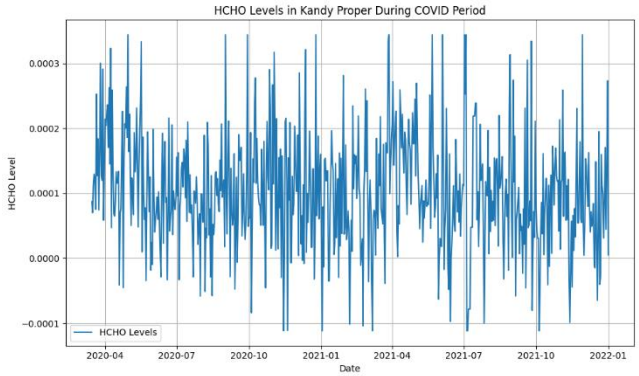
Mean HCHO levels in Bibile, Monaragala:
Pre-COVID: 0.00011646276686692911
During COVID: 0.00011776573626440579
Post-COVID: 0.00012150880129932713




```
Mean HCHO levels in Kurunegala
Proper:
Pre-COVID: 0.0001362629565870312
During COVID: 0.00012451119902587434
Post-COVID: 0.0001252989846131842 3
```




```
Mean HCHO levels in Jaffna Proper:
Pre-COVID: 0.00010733549787751874
During COVID: 0.00010742490492241578
Post-COVID: 0.0001151404834128448 5
```




```
Mean HCHO levels in Kandy Proper:
Pre-COVID: 0.00011808399450647088
During COVID: 0.00010258652691374019
Post-COVID: 0.0001092129507054740 5
```

These graphs present the HCHO levels in various cities from April 2020 to January 2022. And from the side it shows the mean HCHO values in each city before covid period, during covid period, and after the covid period.

4. Machine Learning

The application of machine learning techniques in time-series forecasting provides a robust framework for predicting future HCHO levels based on historical data. This chapter delves into the deployment of ARIMA and SARIMAX models, which are well-suited for time-series data exhibiting non-stationarities and seasonal patterns.

4.1 ARIMA

The Autoregressive Integrated Moving Average (ARIMA) model is a cornerstone in time-series analysis, known for its efficacy in modeling data that shows evidence of non-stationarity. ARIMA models are characterized by three parameters: (p) autoregressive, (d) integrated, and (q) moving average terms. The model was specifically chosen for its ability to model a wide array of time series data including those with trends and non-seasonal fluctuations.

Before doing the forecasting in each city, I checked whether the dataset is stationary using ADF test. The dataset was stationary with an ADF statistic of -12.933626986840022. After training the model below is the future predicted HCHO values for each city.

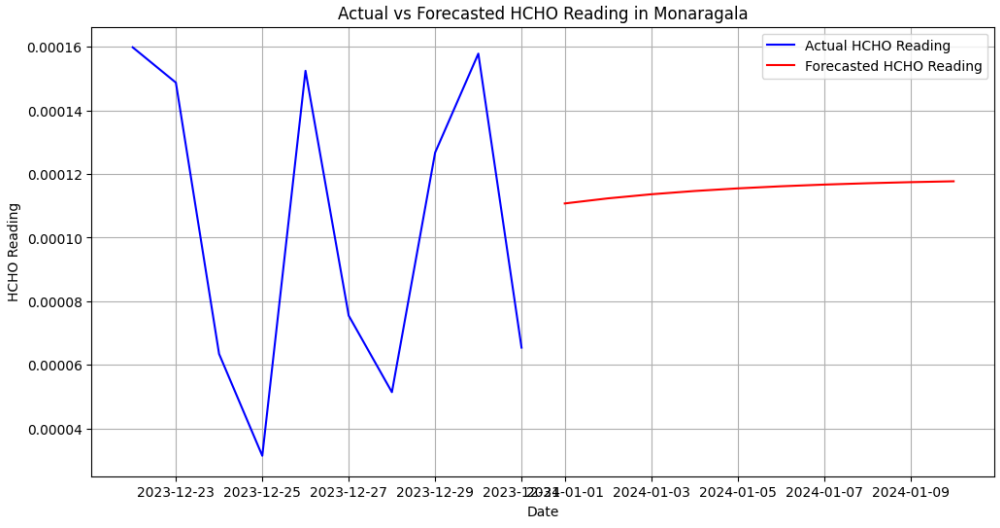


Figure 19: ARIMA Model - Bibile, Monaragala

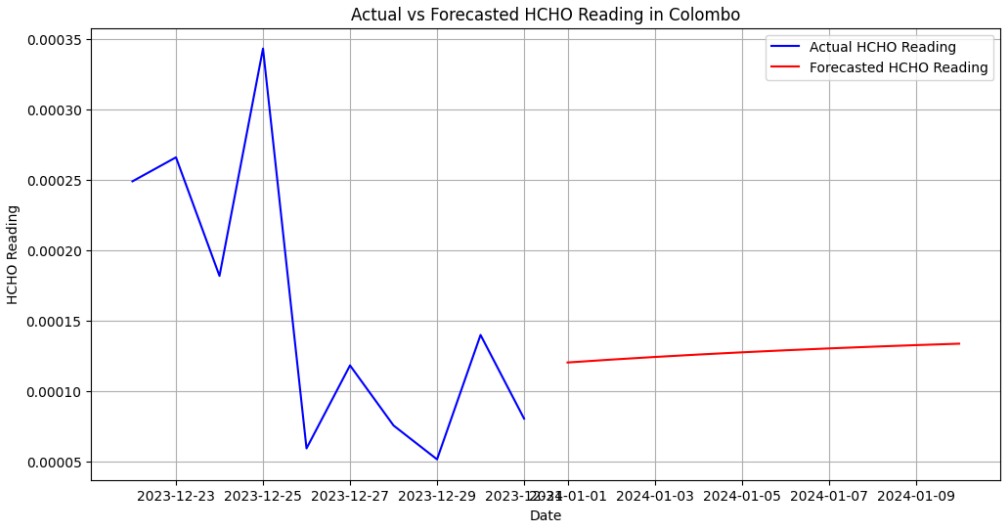


Figure 20: ARIMA Model - Colombo Proper

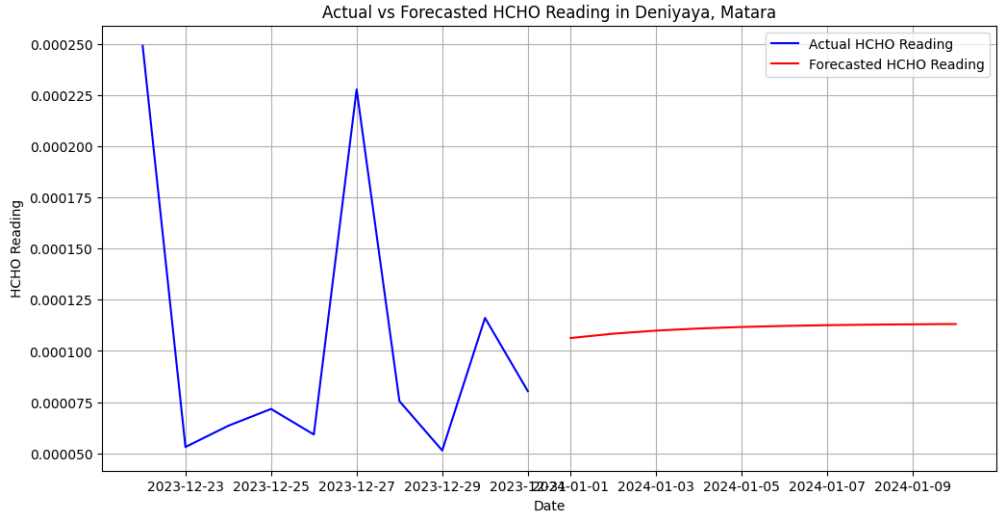


Figure 21: ARIMA Model - Deniyaya, Matara

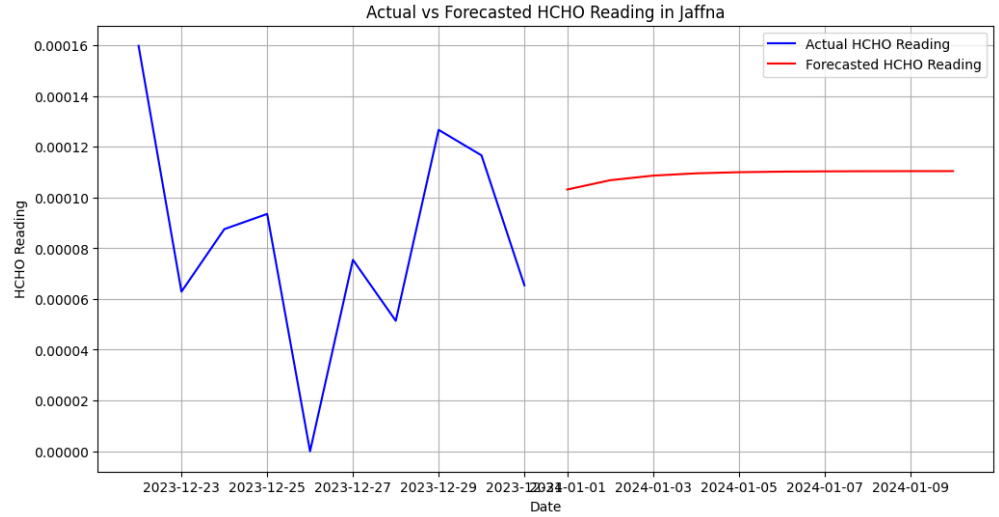


Figure 22: ARIMA Model -Jaffna Proper

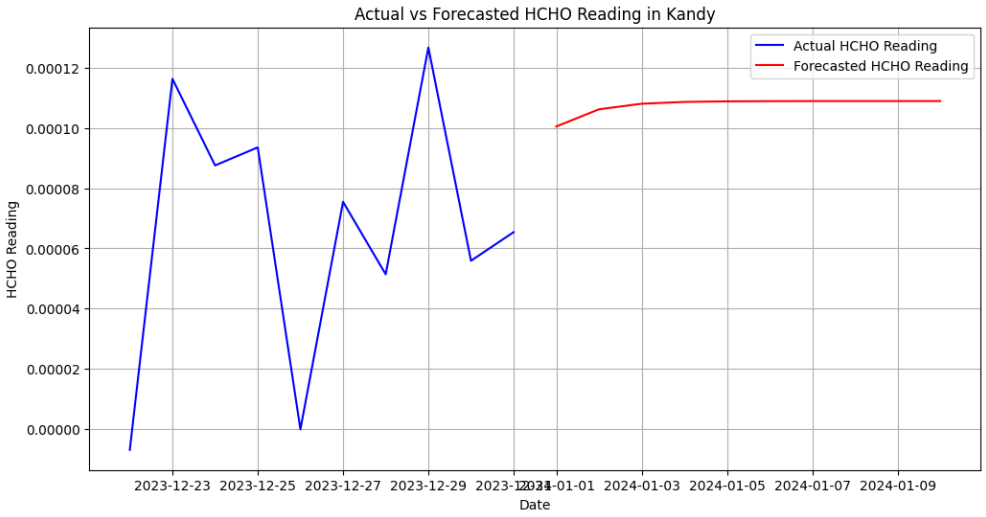


Figure 23: ARIMA Model - Kandy Proper

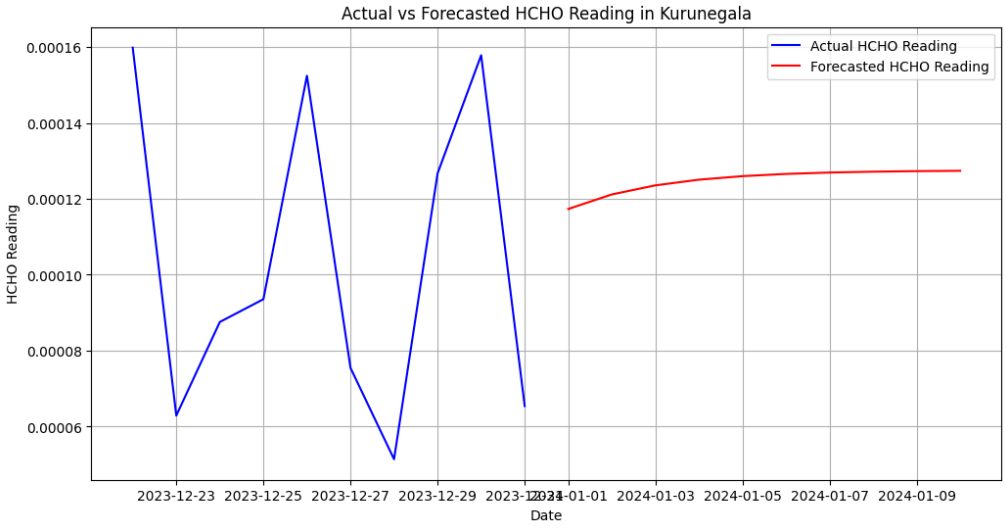


Figure 24: ARIMA Model - Kurunegala Proper

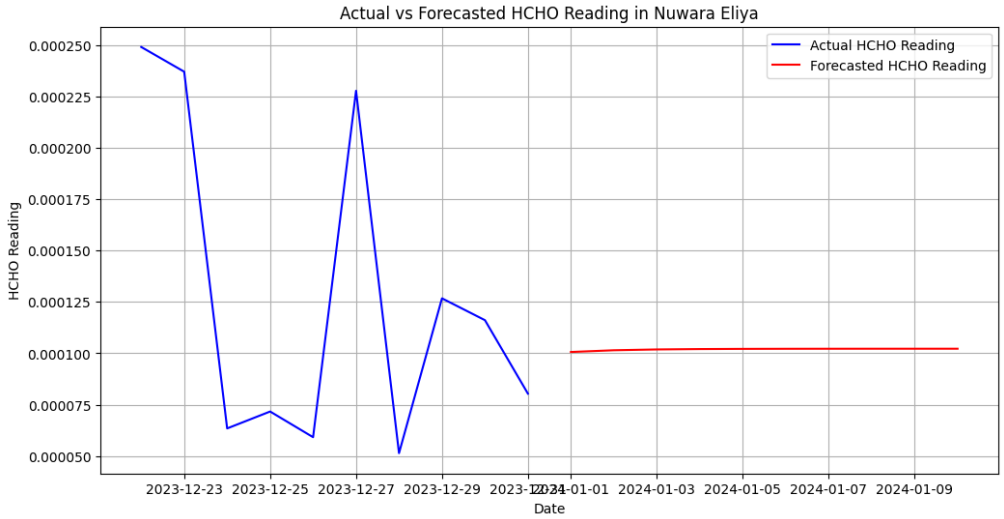
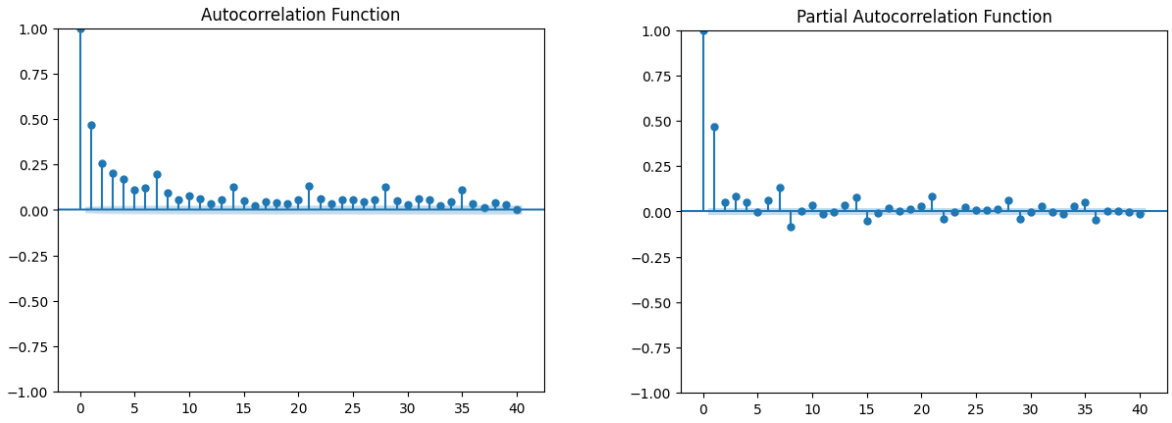


Figure 25: ARIMA Model - Nuwara Eliya Proper

Based on the above graphs the predicted line doesn't show any fluctuations in the values. It doesn't take seasonality into account. Therefore, I chose the SARIMAX Model.

4.2 SARIMAX

Seasonal ARIMA with eXogenous variables (SARIMAX) extends the ARIMA model by incorporating both seasonal components and exogenous variables, allowing for a more nuanced understanding of seasonal effects and external influences on HCHO levels. This model is particularly useful in environmental data analysis where seasonality and external factors such as temperature or industrial activity significantly impact the variables of interest.



Monaragala

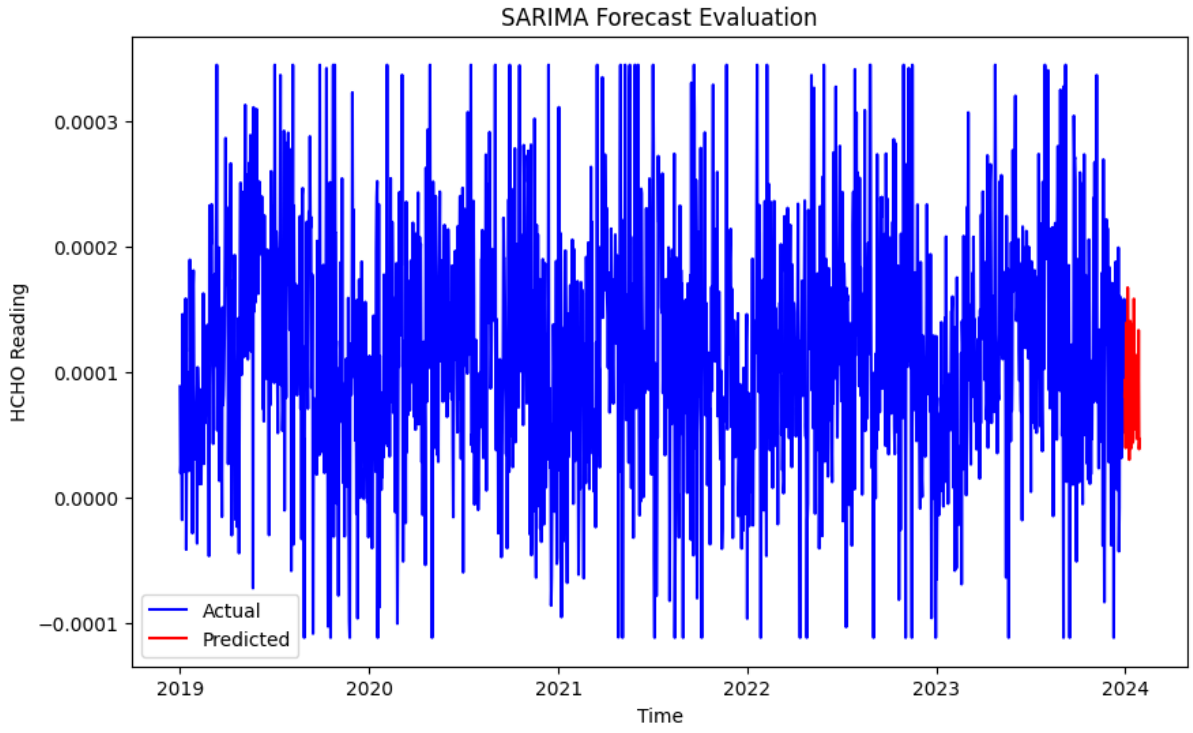


Figure 26: SARIMAX Future Forecast Evaluation - Bibile, Monaragala

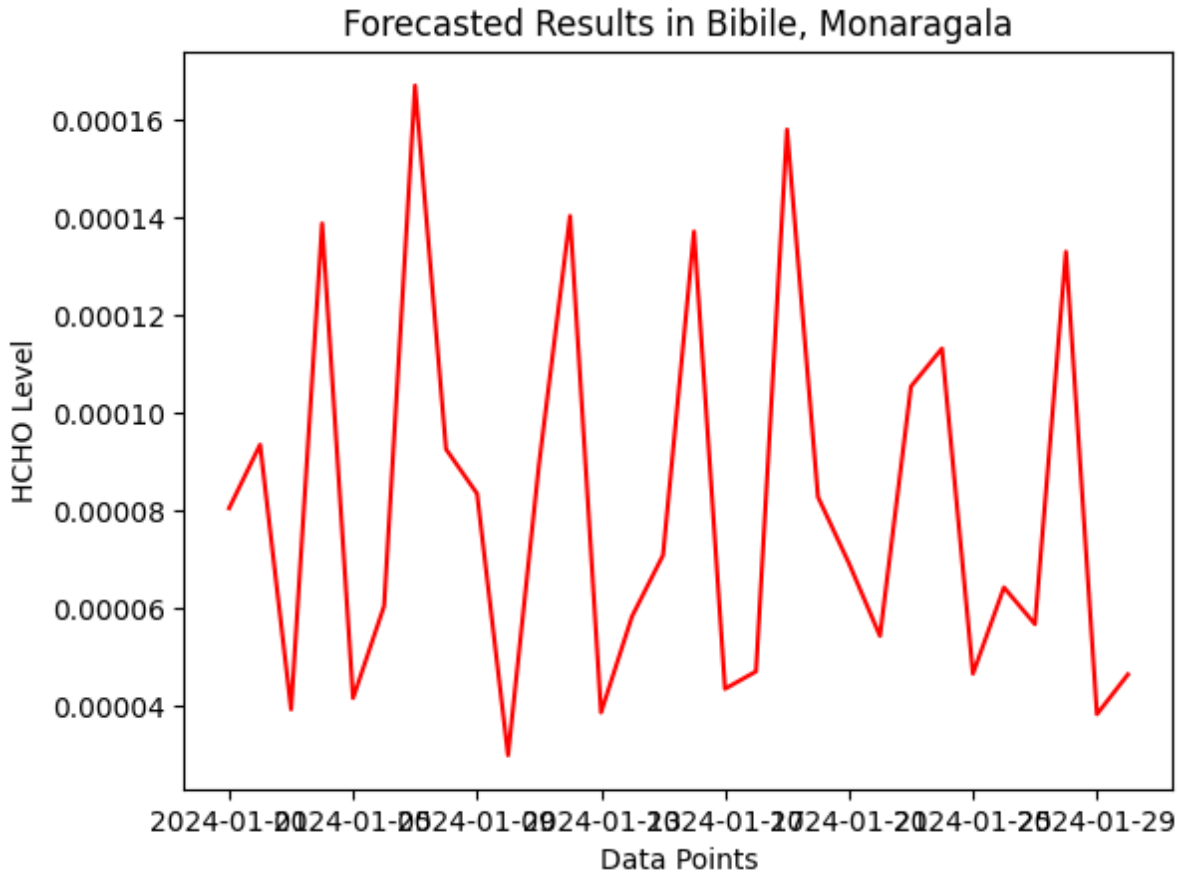


Figure 27: SARIMAX Future Prediction - Bibile, Monaragala

Colombo

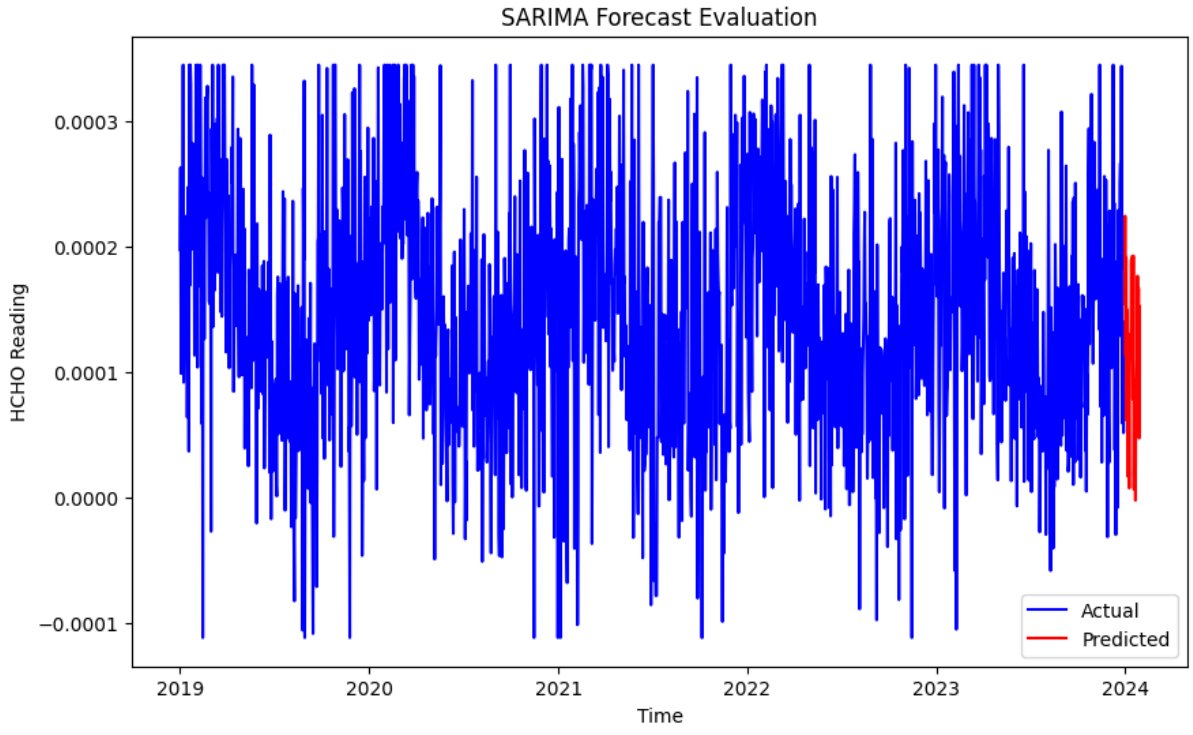


Figure 28: SARIMAX Future Forecast Evaluation - Colombo Proper

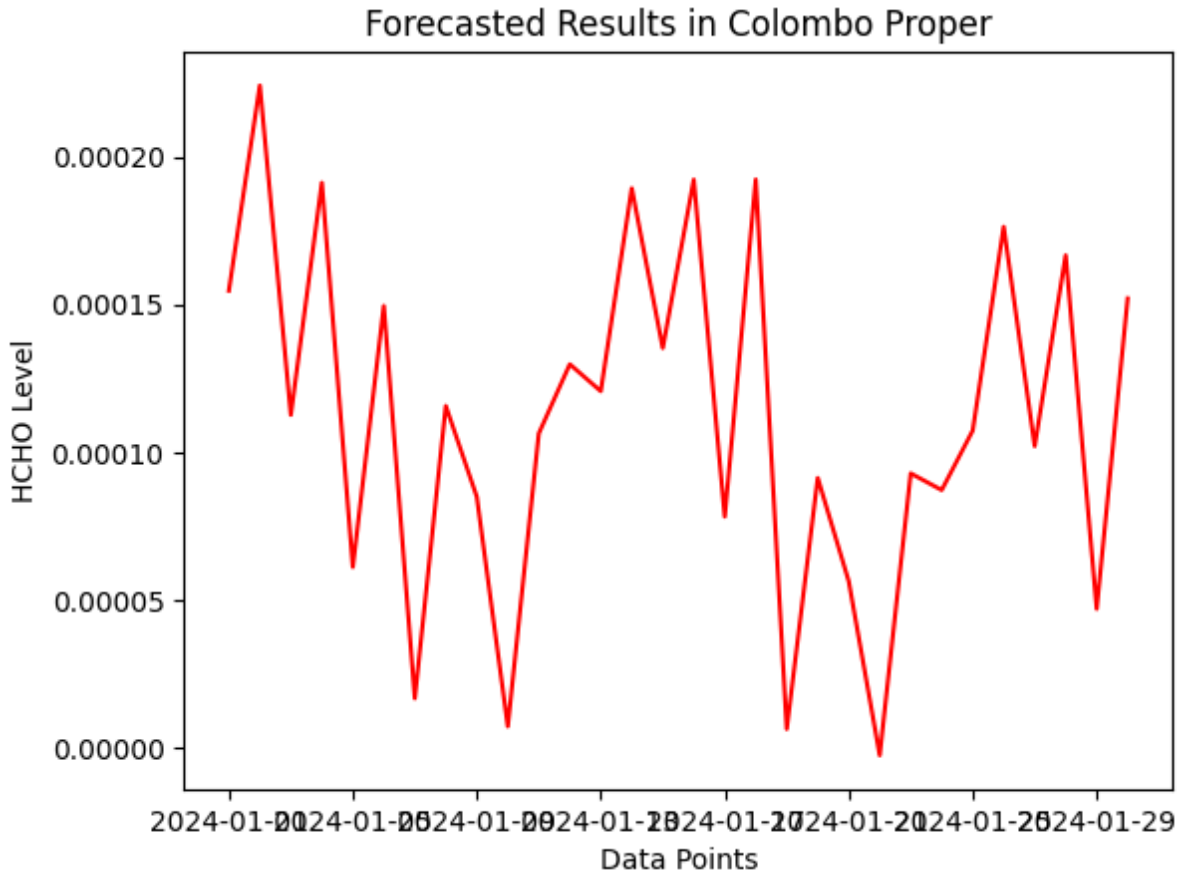


Figure 29: SARIMAX Future Prediction - Colombo Proper

Jaffna

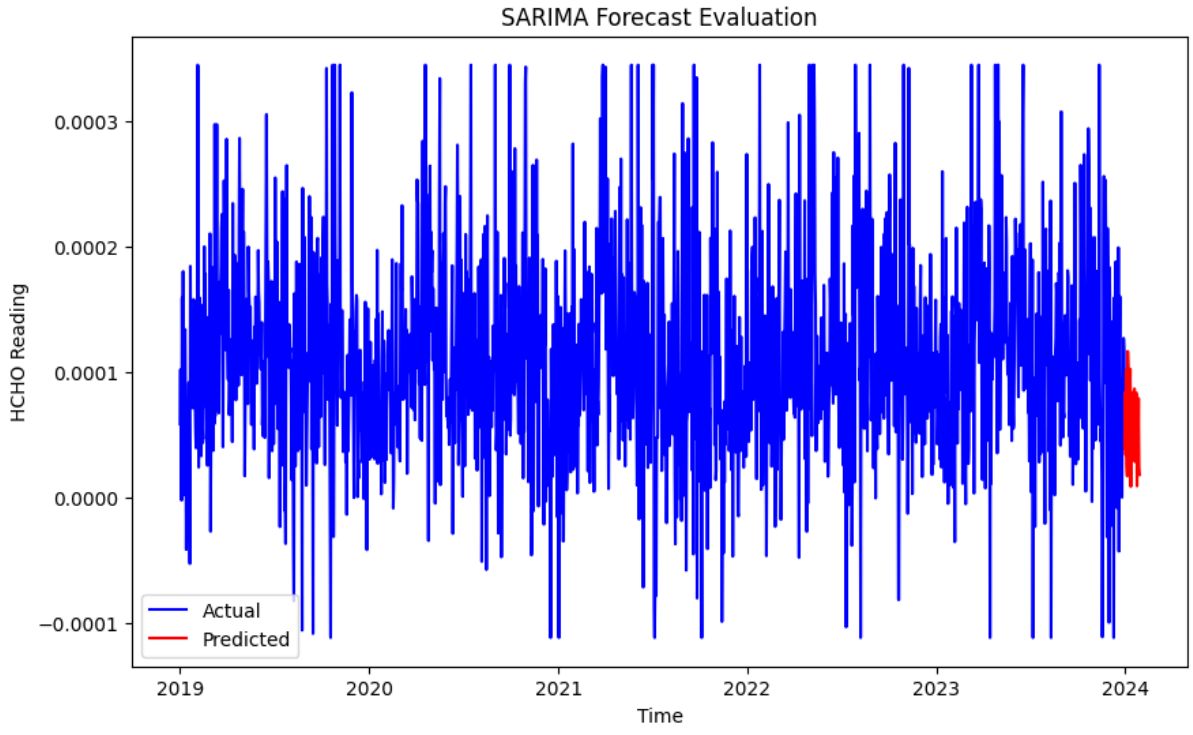


Figure 30: SARIMAX Future Forecast Evaluation - Jaffna Proper

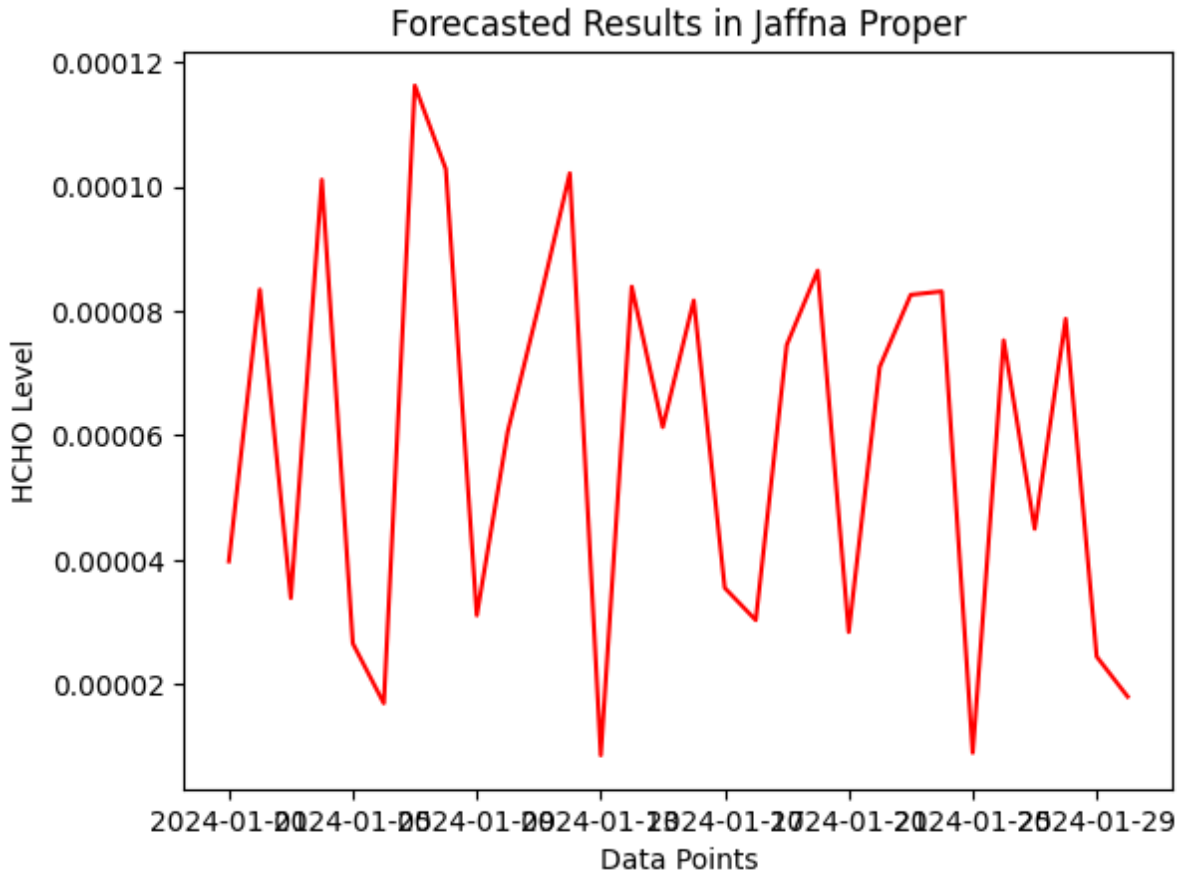


Figure 31: SARIMAX Future Prediction - Jaffna Proper

Matara

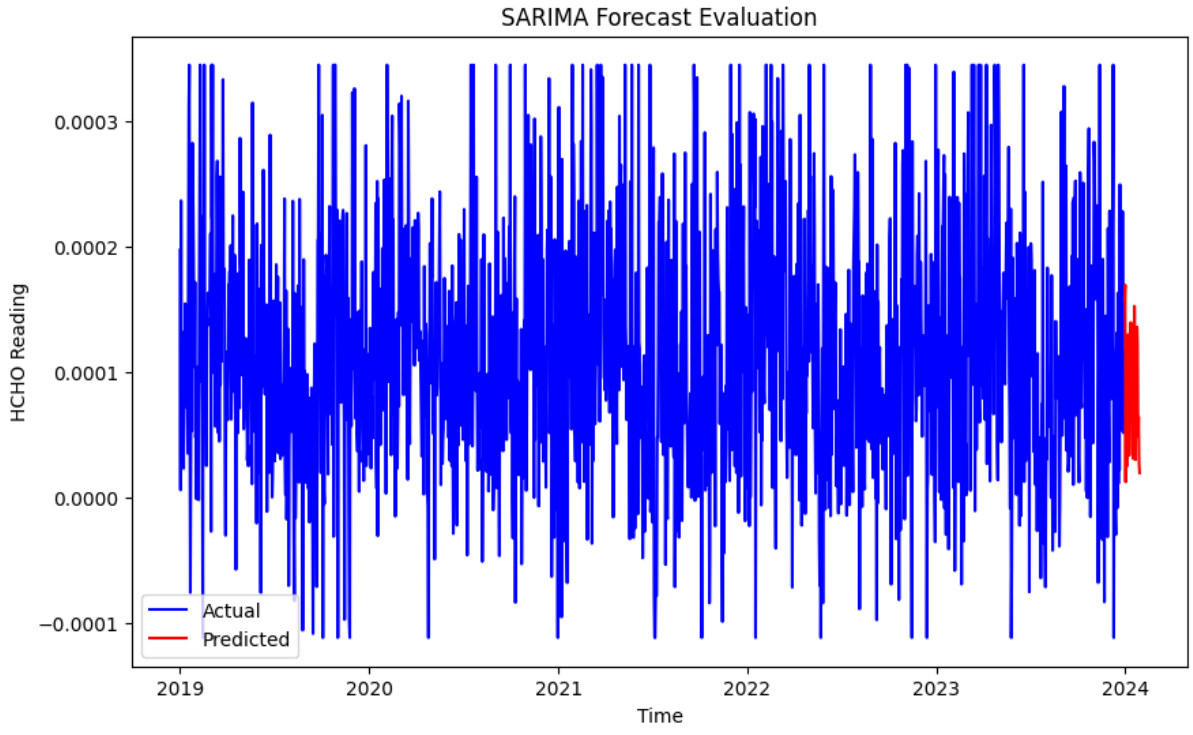


Figure 32: SARIMAX Future Forecast Evaluation - Deniyaya, Matara

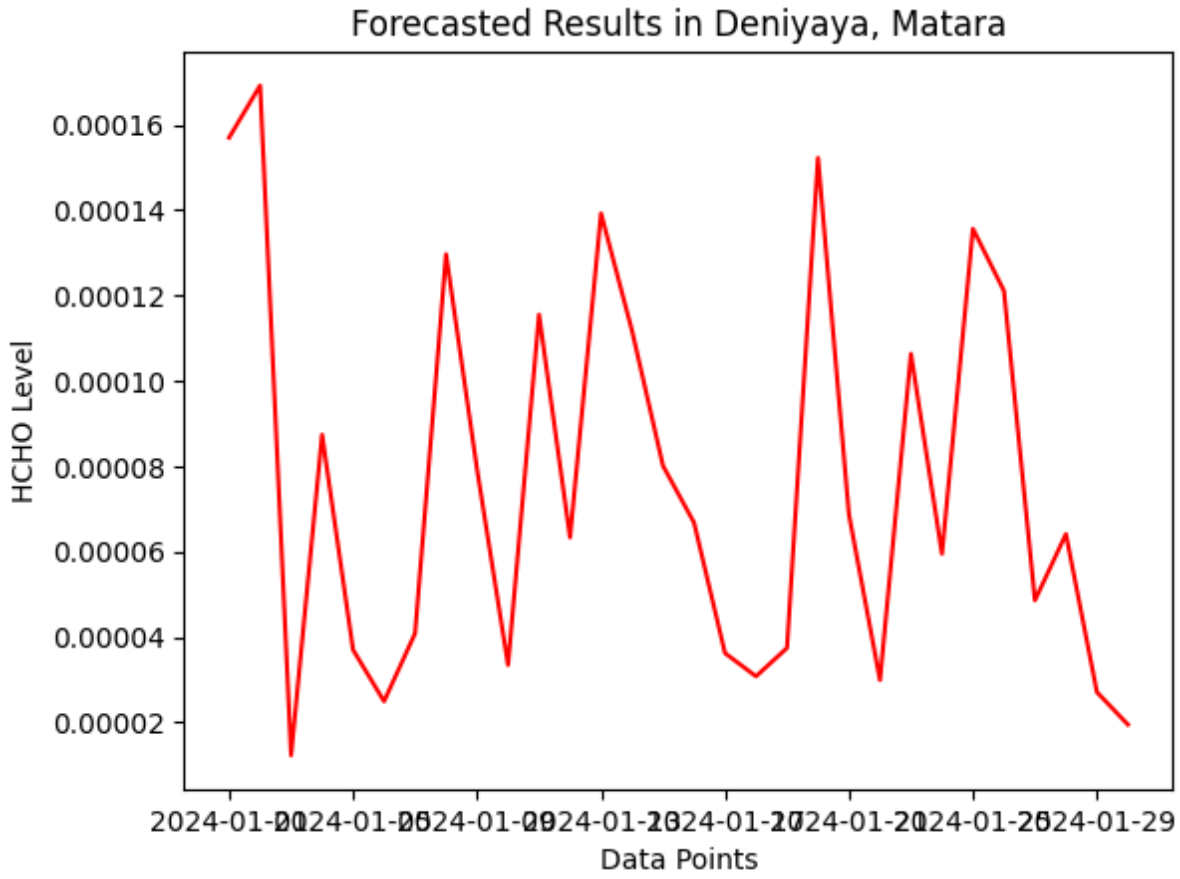


Figure 33: SARIMAX Future Prediction - Deniyaya, Matara

Kandy

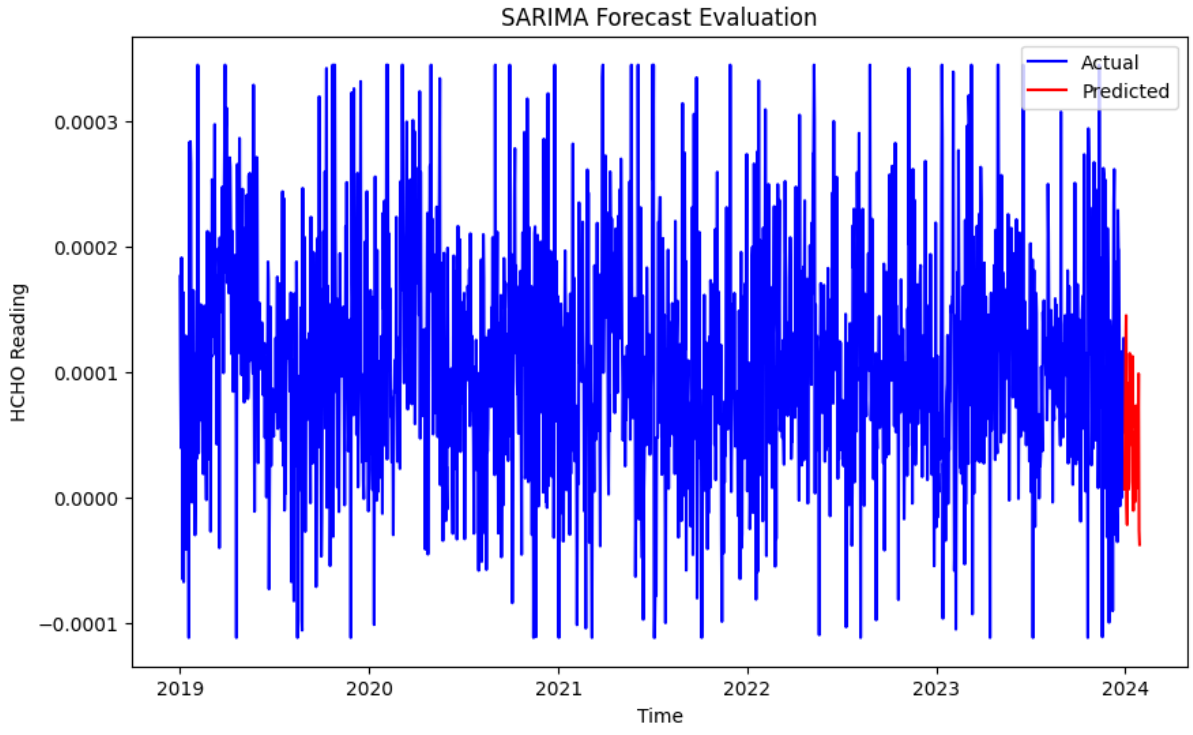


Figure 34: SARIMAX Future Forecast Evaluation - Kandy Proper

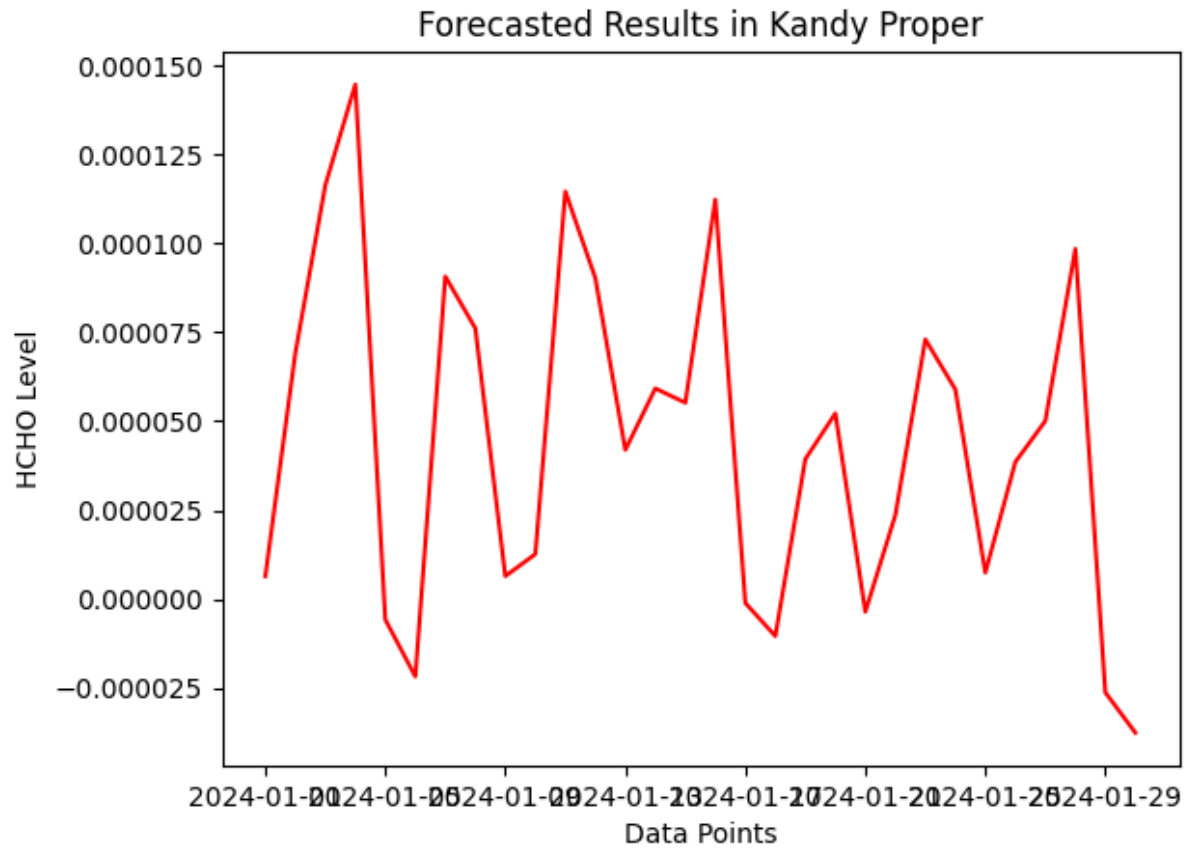


Figure 35: SARIMAX Future Prediction - Kandy Proper

Kurunegala

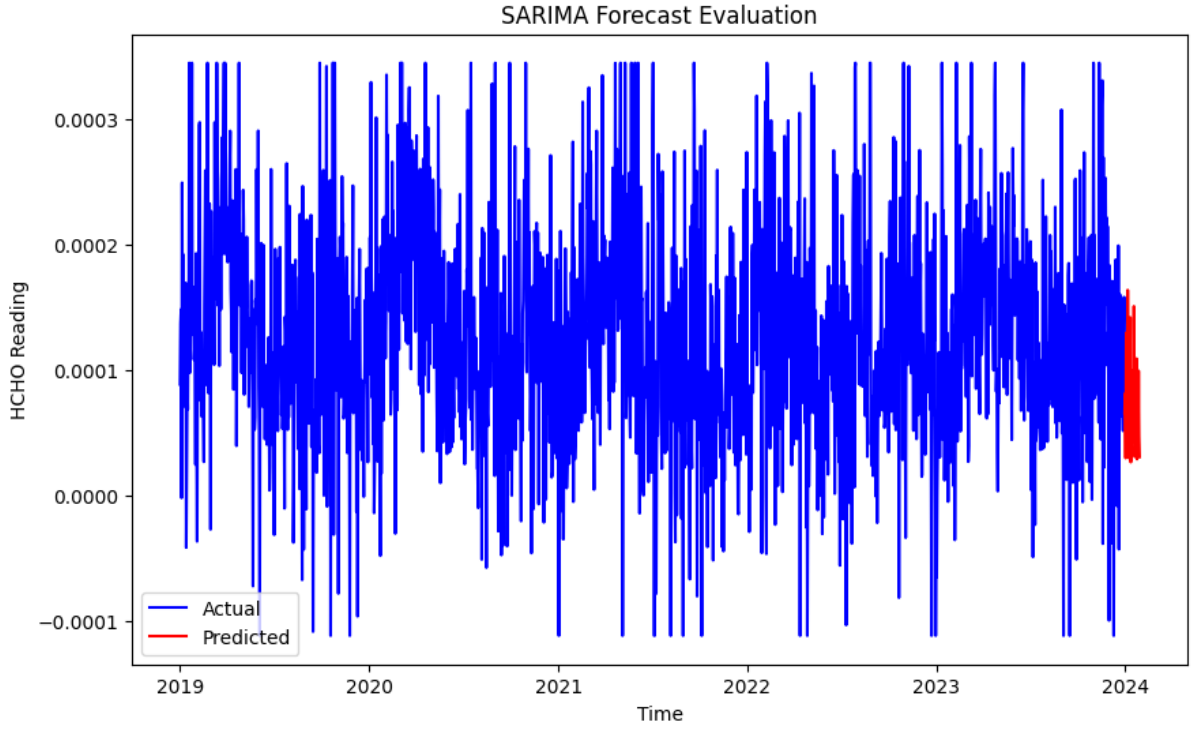


Figure 36: SARIMAX Future Forecast Evaluation - Kurunegala Proper

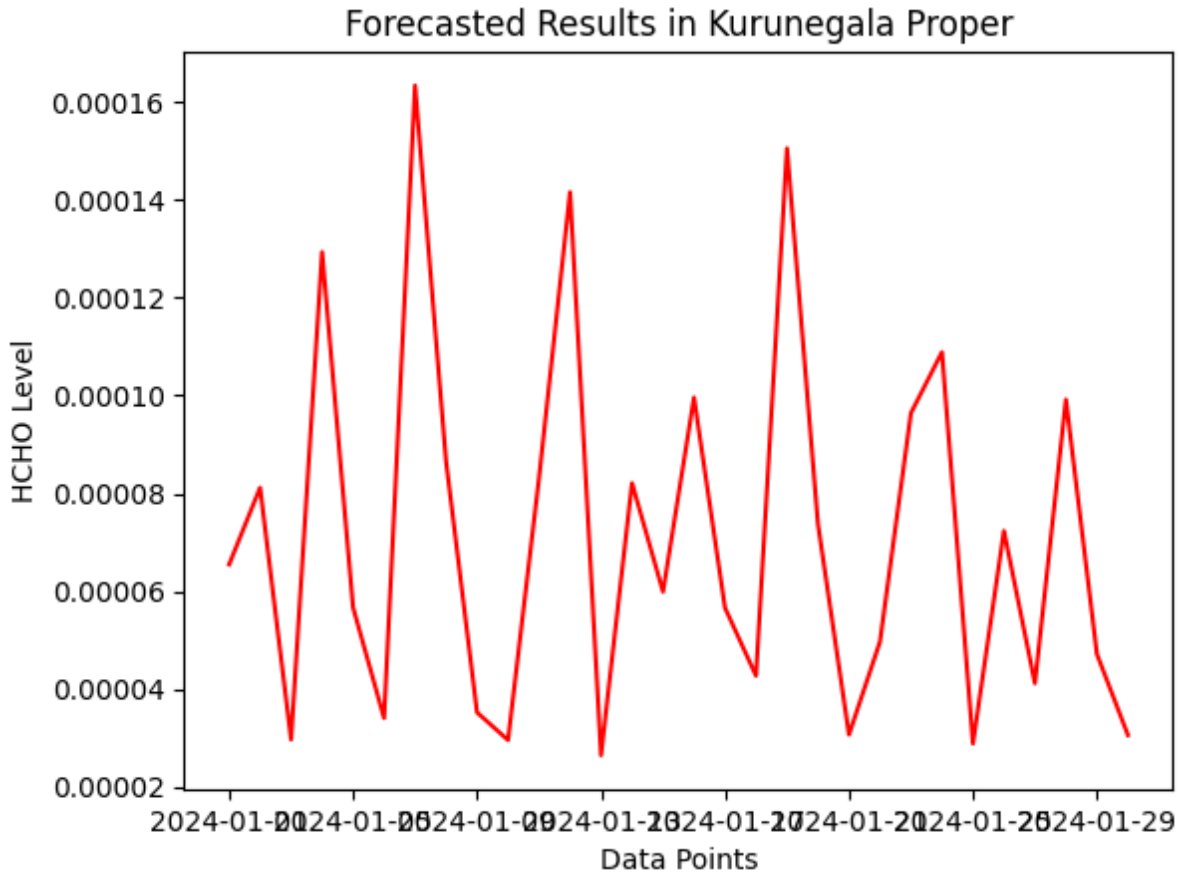


Figure 37: SARIMAX Future Prediction - Kurunegala Proper

Nuwara Eliya

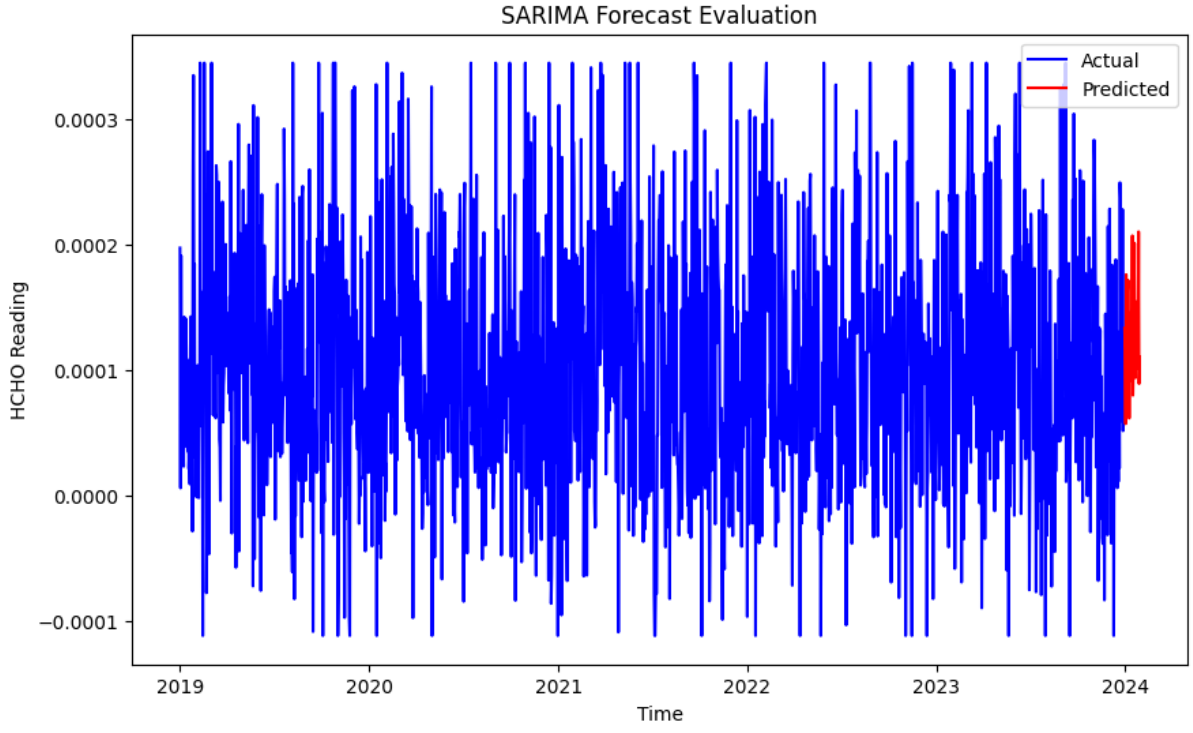


Figure 38: SARIMAX Future Forecast Evaluation - Nuwara Eliya Proper

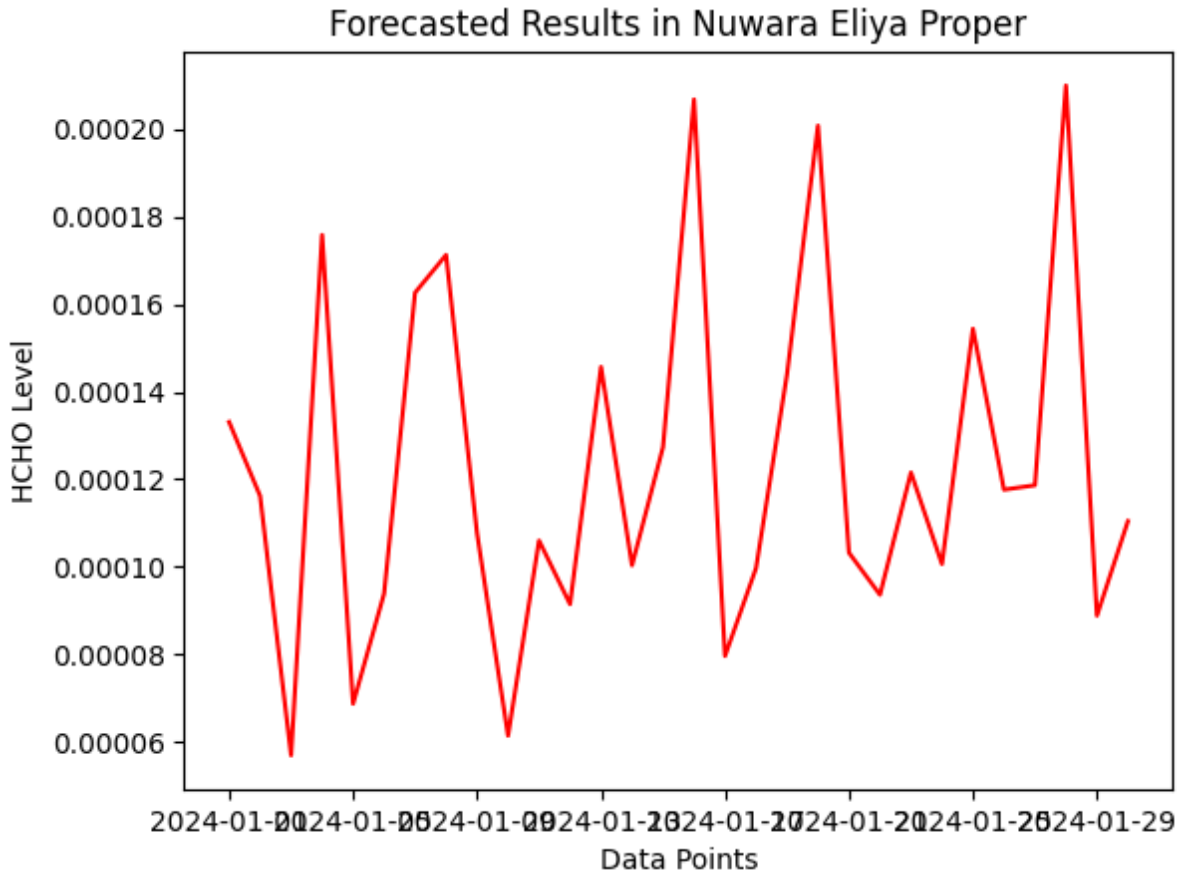


Figure 39: SARIMAX Future Prediction - Nuwara Eliya Proper

4.3 Model Performance

To evaluate the model performance, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) will be used.

Bibile, Monaragala

```
MSE: 1.6796953349658348e-07
RMSE: 0.0004098408636246311 4
MAE: 0.00036051265156532004
```

Colombo Proper

```
MSE: 1.3337234371025262e-07
RMSE: 0.00036520178492205184
MAE: 0.0003142697056609586
```

Jaffna Proper

```
MSE: 1.865225680220123e-08
RMSE: 0.00013657326532744697
MAE: 0.00011198352097043276
```

Deniyaya, Matara

```
MSE: 2.0870677364679247e-07
RMSE: 0.00045684436479701977
MAE: 0.00040484597903267473
```

Kandy Proper

```
MSE: 6.092272602735443e-07
RMSE: 0.0007805301149049563
MAE: 0.0006920063822271161
```

Kurunegala

```
MSE: 1.0702854698598313e-08
RMSE: 0.00010345460211415591
MAE: 8.277757393169253e-05
```

Nuwara Eliya Proper

```
MSE: 4.927961291291571e-08
RMSE: 0.00022199011895333474
MAE: 0.00019056843662669615
```

According to the above MSE, RMSE, MAE values, we can identify that the model is accurate, the predicted values have a close fit to the actual data, and the model is a best fit for prediction. as those values are low.

4.4 Limitations

4.4.1 Limitations of ARIMA Models

- **Stationarity Requirement**: ARIMA models require the data to be stationary, meaning the statistical properties such as mean, variance, and autocorrelation are constant over time. This often necessitates transformations like differencing, which can complicate the model and potentially lose some information.
- **Linearity**: ARIMA models assume a linear relationship between lagged observations and the forecasted values. This assumption may not hold true in cases where the underlying processes are inherently non-linear, as often seen in complex natural phenomena or financial markets.
- **Predictive Power Limited to Short-term Forecasts**: ARIMA models are generally better for short-term forecasting. As the forecasting horizon increases, the accuracy of ARIMA predictions tends to deteriorate rapidly, particularly in volatile or highly dynamic environments.

4.4.2 Limitations of SARIMAX Models

- **Increased Complexity**: SARIMAX models extend ARIMA by incorporating seasonal components and exogenous variables, which increases the model's complexity. This complexity requires careful calibration of more parameters, including seasonal differencing terms and the integration of external variables, making the model more challenging to optimize and interpret.
- **Sensitivity to Specified Exogenous Variables**: The performance of SARIMAX models heavily depends on the choice and relevance of the exogenous variables included. Incorrectly specifying these variables or using poorly correlated external data can lead to misleading results and poor forecasts.

4.5 Potential Improvements

- **Volatility Modeling**: Integration with GARCH Models: For time series data with apparent volatility clustering, such as financial markets, combining ARIMA/SARIMA with Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models can effectively model and predict the level of volatility.
- **Advanced Machine Learning Models**: Incorporating machine learning models such as Random Forests or Gradient Boosting Machines could capture non-linear interactions that are not possible to model with SARIMAX or ARIMA.
- **Deep Learning Approaches**: Implementing deep learning models like LSTM (Long Short-Term Memory networks) or GRU (Gated Recurrent Units) which are well-suited for sequence prediction tasks and can handle very long sequences with complex patterns.

5. Future Enhancements

- **Expanding Data Sources**: Integrating additional data sources, such as direct emission measurements from industrial activities and more detailed meteorological data, might provide a more comprehensive understanding of the factors influencing HCHO levels.
- **Model Ensemble Techniques**: Combine multiple predictive models to enhance forecast reliability and accuracy. It will help to reduce model bias and variance, leading to more robust predictions.
- **Develop a Real-time Monitoring System**: Implement a system that can process and analyze data in real-time to provide up-to-date information on HCHO levels.
- **Scalability and Cloud Computing**: Utilize cloud computing for data analysis scalability, supporting larger datasets and more complex models.

6. Similar Studies

There are many researches carried out on the world for analysis of HCHO gas distribution for making desertions in several domains. Zhu et al. (2017) conducted an extensive analysis of long-term trends in formaldehyde (HCHO) columns utilizing data from the Ozone Monitoring Instrument over a decade (2005-2014). The study revealed significant spatial variations in HCHO levels across North America, attributed to both anthropogenic activities and natural sources. Key findings included marked declines in HCHO concentrations in regions with stringent air quality regulations and increased emissions in areas experiencing industrial growth, such as the Cold Lake Oil Sands. Their research demonstrates the critical role of satellite-derived atmospheric data in assessing the impact of environmental policies and economic activities on air quality. This study serves as a pivotal reference for our analysis, illustrating the application of remote sensing technology in environmental monitoring and its potential for informing policy-making.

7. References

Zhu, L. et al. (2017) 'Long-term (2005–2014) trends in formaldehyde (HCHO) columns across North America as seen by the OMI satellite instrument: Evidence of changing emissions of volatile organic compounds', Geophysical Research Letters, 44(13), pp. 7079–7086. https://doi.org/10.1002/2017GL073859.

GitHub link to the project - https://github.com/HJayashan/HCHO-Prediction.git