FORECASTING THE PARTICULATE MATTER 2.5: A CASE STUDY IN CHALOEM PHRA KIAT DISTRICT, SARABURI PROVINCE, THAILAND

Thanatorn Chuenyindee[1] Puthipong Tanaveerakul[2*] Ardvin Kester S. Ong[3]
Chatwaleerat Sakulsuksomboon[4]

Received : February 14,2025
Revised : August 4, 2025
Accepted : August 27, 2025

Abstract

The airborne particulate matter, which has been especially concerned about the case of PM2.5, has been a prominent factor heavily affecting everyday lives of Chaloem Phra Kiat District's residents. It has negatively influenced the health of the residents, especially as it has become evident in allergic-related skin conditions among the residents. Having understood the decline in the quality of life because of PM2.5 pollution, this study has sought to make predictions for the amounts of PM2.5 concentrations based on the previous data over the last 4 months, 6 months, and 1 year. The forecasting models used in the analysis revolved around the Auto-Regressive Integrated Moving Average (ARIMA), Vector Auto Regression (VAR), and Long Short-Term Memory (LSTM). The Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) were used to estimate the accuracy and performance of these models employing MATLAB and Orange as the data analysis software. It was found that the error rates between the two models of ARIMA and VAR were comparable, while the LSTM model showed significantly lower error rates, with the lowest MAE of 1.67 μg/m³ and MAPE of 7.94%, which was indicative of a better capacity to forecast. Additionally, the study showed clear seasonal fluctuations in the PM2.5 concentrations, which grew steadily to peak during the winter, then fell in summer, and finally fell to their lowest during the rainy season. For example, the peak monthly average in January reached over 55 μg/m³, while in August, it dropped below 15 μg/m³. A consistent cyclical pattern was found every year. As a benchmark forecasting and comparative analysis, this research laid a foundation for further research studies, possibly using advanced machine learning algorithms for further improvement of predictive accuracy and robustness of the models involved.

Keywords: Air quality analysis, ARIMA, LSTM, PM2.5 forecasting, Seasonal variation, VAR

---

[1] *Department of Industrial Engineering and Aviation management, Division of Education, Navaminda Kasatriyadhiraj Royal Air Force Academy, e-mail: thanatorn_chu@rtaf.mi.th*

[2] *Faculty of Business Administrator, University of the Thai Chamber of Commerce, e-mail: puthipong_tan@utcc.ac.th*

[3] *School of Industrial Engineering and Engineering Management, Mapúa University, e-mail: aksong@mapua.edu.ph*

[4] *Faculty Of Management Science, Phetchaburi Rajabhat University, e-mail: Chatwaleerat.sakul@gmail.com*

** Corresponding author, e-mail: puthipong_tan@utcc.ac.th*

Introduction

Air pollution is an acute environmental concern within a local and global level instigated primarily by the actions of man such as open-burning, transport, electrical generation, industrial activities, and atmospheric pollutants were we to damage the environment (World Health Organization, 2024). Air pollution was a recognized big issue of environmental health, around 4.2 million people died due to air pollution early in 2019 (World Health Organization, 2024). Fine particulate matter, also known by the World Health Organization as carcinogenic, with a size of 2.5 microns or less is an alleged health disaster (Greenpeace Southeast Asia, 2024). Industrial activities are the main PM2.5 pollution producers in the Chaloem Phra Kiat District, Saraburi Province, and this is associated with serious complications in both environment and public health (Pollution Control Department of Thailand, 2020).

In recognition of these important factors, global intensification has been done to fight air pollution, namely focusing on PM2.5 volumes in regard to the fact that these volumes are microscopic with just about 20 times smaller than the human hair diameter. This size makes them invisible and able to penetrate around conventional face masks and the normal nasal filtering systems, exposing and thus contributing to more health dangers (United States Environmental Protection Agency, 2024). Alongside the respiratory impacts, the lives of Chaloem Phra Kiat residents are affected adversely on a day-to-day basis due to PM2.5, evident in its dermatological effects. Sensations comprehensively observed by patients with anaphylaxis are hives, itching, and severe exacerbations to the sensitive regions like eyes, mouth, nose, face, groin, and joints, leading to cell-level skin injuries (Sarla, 2020).

New studies are stressed to highlight considerably distinct health effects related to the influence of PM2.5, as the connection to the rising morbidity and mortality rates because of the adverse respiratory and cardiovascular junctionary systems (Garcia et al., 2023). A systematic

review by Li et al. (2022) further elaborates the organ specific damage in which genotoxicity, oxidative stress, as well as infections, inflammatory responses besides reported damage to renal, gastrointestinal, reproductive and neurological systems are cited. Related concerns over

exposure to long-term PM2.5 exposure risks to brain issues highlight the dire need for proactive intervention, given current ambiguities surrounding how the brain can recover and be restored over the long term.

Acknowledging the further decline in quality of life that can be attributed to the PM2.5 pollution, researchers call attention to the importance of accurate forecasting models so that effective public healthcare can prevent the current situation. Furthermore, PM2.5 poses a major obstacle to the visibility, increasing the hazard of traffic-related accidents (Hyslop, 2009). In this study, the MATLAB and Orange software platform was used, this is a high-class-versatile visual programming tool that combines machine learning and high-level-data analysis techniques. In particular, the Auto Regressive Integrated Moving Average (ARIMA) and Vector Auto Regression (VAR) models were applied as they are well-verified models for the forecast of concentrations of PM2.5 (de Myttenaere et al., 2016). Prestations of models and exactitude of forecasts were intensely tested by using Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

As core objectives of this research, it had been necessary to identify the best forecasting ARIMA and VAR models adapted to forecasting PM2.5 concentration in particular within Chaloem Phra Kiat District, Saraburi Province and conducting predictive trends in future to understand practical measures for mitigating these future trends. The implication of this research is to give critical perspectives for improving the preparedness of communities, guiding programs in public health management, and assisting the officials of the government with programs of mitigating specific policies. Moreover, these findings can provide meaningful potential for benchmarking, which could be successfully modified and implemented by the neighboring regions and international communities that also encounter environmental challenges related to water security.

Objectives

1. To develop and evaluate forecasting models (specifically ARIMA, VAR, and LSTM) for predicting PM2.5 concentration level in Chaloem Phra Kiat District using historical air quality data.

2. To compare the forecasting accuracy of these models across different time horizons (4 months, 6 months, and 1 year) using statistical error metrics such as MAE, MAPE, and RMSE.

3. To analyze seasonal patterns and trends in PM2.5 concentrations, with the aim of informing public health interventions and local environmental management policies.

Materials and Methods

1. Data Collection

This study received data support from the Pollution Control Department, the Ministry of Natural Resources and Environment. The researcher nevertheless obtained air pollution data, in particular the daily average of PM2.5 (particulate matter with a diameter of 2.5 microns or less) data over the past year from the monitoring station at the Chaloem Phra Kiat Police Station in Saraburi Province, Thailand.

2. Data Analysis

The PM2.5 concentration data from the past two years was converted into a .csv file format for integration into the Orange software program and MATLAB software program. Subsequently, the PM2.5 concentration data were analyzed using the ARIMA, VAR, and LSTM forecasting models to inform future predictions of PM2.5 levels. The analysis was conducted with a focus on different periods to test the accuracy of the forecasts. The steps for data analysis and the PM2.5 concentration data were divided into three sets:

- Daily average PM2.5 concentrations over one year (June 2022 - December 2022).
- Daily average PM2.5 concentrations over six-months (December 2022 - May 2023).
- Daily average PM2.5 concentrations over four-months (February 2023 - May 2023).

Afterward, the daily average PM2.5 concentration data were formatted as a time series. Dates were organized sequentially to facilitate their use in time series functions for subsequent forecasting, as illustrated in Figure 1.

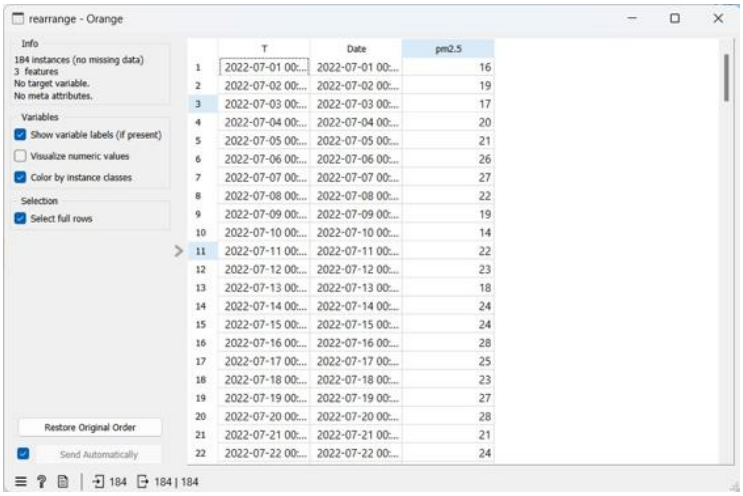
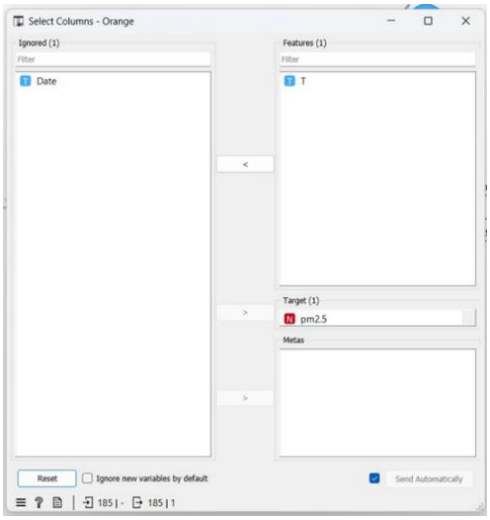


Figure 1 Time series windows

The PM2.5 concentration variable was entered into the Target field to display the results in the graph following the forecast, as shown in Figure 2a. The parameters p, d, and q of the ARIMA model were then adjusted to optimize the model using a grid search method, establishing a range from 0 to 2 in order to achieve the lowest possible error, as depicted in Figure 2b.



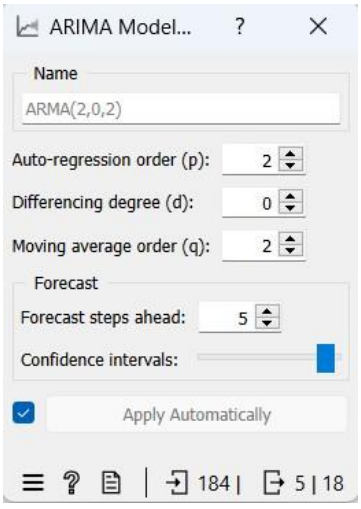


Figure 2a Columns selection windows Figure 2b ARIMA model parameter adjustment

The parameters of the VAR model were adjusted to optimize the model by minimizing the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), as illustrated in Figure 3.

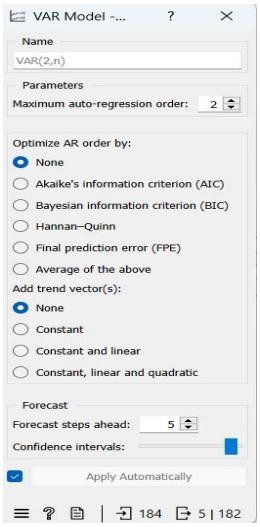


Figure 3 VAR model parameter adjustment

3. Evaluation and Comparison of Data Analysis Results

The purpose of this evaluation and comparison of the results of data analysis is to judge the accuracy and reasonability of time frames for both forecasting models. This was carried out using the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), and the Mean Absolute Percentage Error (MAPE) to assess the output quality of the predictions. Lower error values indicate better performance on the employees' tests. Low numbers are indicative of higher effectiveness in the case of MAE; a value of zero is a high value of accuracy of the model; the model predicts the value of y with 100% accuracy. Concerning MAPE, a lower percentage of less than 10% is highly accurate, 10-20% is good, 20-50% is acceptable, and figures exceeding 50% are low in terms of accuracy of forecasts.

Each forecasting method offers distinct advantages and limitations. The ARIMA model is widely used for univariate time series forecasting due to its simplicity and interpretability; it performs well when data shows linearity and autocorrelation but may not capture nonlinear patterns effectively (de Myttenaere et al., 2016). The VAR model is useful for multivariate time series with interdependent variables, assuming linear relationships among them, although it requires stationarity and may underperform when nonlinear or long-memory processes are present (Wang et al., 2017). In contrast, the LSTM model, a type of recurrent neural network, can learn complex and nonlinear temporal dependencies, which makes it highly effective for

long-sequence forecasting problems. However, it requires a large amount of training data, significant computational power, and careful hyperparameter tuning to avoid overfitting (Li et al., 2022; Li et al., 2023).

Results and discussion

In the pursuit of identifying the most effective model for forecasting PM2.5 concentrations, the ranges for the parameters p, d, and q to verify whether they fall within the desired testing intervals were established. Subsequently, models were constructed based on the defined parameter sets to compare their performance by evaluating the error values across each domain. The model that provided the most suitable results were selected using datasets for forecasting over one year, six months, and four months. The optimal ARIMA model was determined through a grid search method, alongside the VAR model. The average error metrics MAE and MAPE indicated that appropriate models could be chosen based on parameter adjustment tests as presented in Table 1, Table 2, and Table 3.

From Table 1, the ARIMA model testing using the grid search method within the range of 0 to 2 reveals that the ARIMA(1,0,0) model has MAE = 4.89, MAPE = 0.185 and RMSE = 15.0, which are the lowest values found. The visual representation of ARIMA forecasting is depicted in Figure 4 and VAR in Figure 5.

The ARIMA model was calibrated using a grid search method with parameter ranges set from 0 to 2. Among the tested configurations, ARIMA(1,0,0) demonstrated the best performance, yielding the lowest MAE of 4.89 and MAPE of 0.185, with an RMSE of 15.0. These values indicate a relatively high accuracy in short- to medium-term forecasting. The forecasted PM2.5 value was approximately 18.975. The forecasting trend illustrated by ARIMA closely aligned with the historical seasonal pattern, capturing both the rising and declining phases of PM2.5 concentrations across the year.

Table 1 Optimal ARIMA model testing (1 Year data)

Data	MAE	MAPE	Forecast	RMSEA
ARIMA(0,0,0)	7.015	0.277	36.663	15.2
ARIMA(0,0,1)	6.56	0.258	26.9301	15.1
ARIMA(0,0,2)	6.005	0.231	20.3322	15.4
ARIMA(0,1,0)	6	0.208	16	15.0
ARIMA(0,1,1)	5.948	0.209	16.0316	15.4
ARIMA(0,1,2)	5.38	0.203	17.7017	15.8
ARIMA(0,2,0)	12.5	0.441	7	15.4
ARIMA(0,2,1)	5.989	0.209	15.9628	15.2
ARIMA(0,2,2)	5.977	0.21	15.9615	15.0
ARIMA(1,0,0)	4.89	0.185	18.975	15.0
ARIMA(1,0,1)	5.019	0.189	18.5008	14.7
ARIMA(1,0,2)	5.035	0.191	18.6736	14.8
ARIMA(1,1,0)	5.963	0.208	16.0229	14.6
ARIMA(1,1,1)	7.027	0.197	20.2466	15.4
ARIMA(1,1,2)	7.286	0.206	19.6389	15.2
ARIMA(1,2,0)	11	0.354	9.16898	15.3
ARIMA(1,2,1)	5.981	0.21	15.9624	14.7

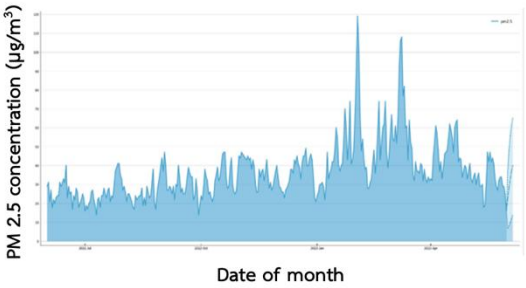


Figure 4 ARIMA Forecasting (1 year)

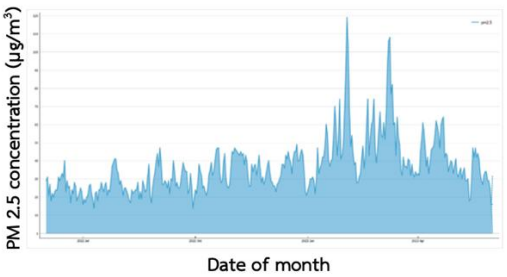


Figure 5 VAR Forecasting (1 year)

From Table 2, the ARIMA model testing using the grid search method within the same range indicates that the ARIMA(0,1,2) model has MAE = 5.48, MAPE = 0.206, and RMSE = 14.3, which are the lowest values found. The visual representation of ARIMA forecasting is depicted in Figure 6 and VAR in Figure 7.

The ARIMA model was again optimized using a grid search within the parameter range of 0 to 2. The best-performing configuration for the 6-month dataset was ARIMA(0,1,2), which produced the lowest MAE of 5.48 and a MAPE of 0.206. The RMSE was calculated at 14.3. This model provided relatively accurate short-term forecasts and captured seasonal fluctuations in PM2.5 levels, with forecasted values closely reflecting the historical downward trend typically observed between December and May.

Table 2 Optimal ARIMA model testing (6-month data)

Data	MAE	MAPE	Forecast	RMSEA
ARIMA(0,0,0)	12	0.292	43.9945	14.4
ARIMA(0,0,1)	9.818	0.269	31.4644	14.0
ARIMA(0,0,2)	8.265	0.234	23.3093	14.1
ARIMA(0,1,0)	6	0.208	16	14.4
ARIMA(0,1,1)	5.869	0.214	15.5982	14.3
ARIMA(0,1,2)	5.48	0.206	17.1702	14.3
ARIMA(0,2,0)	12.5	0.441	7	14.0
ARIMA(0,2,1)	5.998	0.21	15.9001	13.9
ARIMA(0,2,2)	6.098	0.218	15.4409	13.6
ARIMA(1,0,0)	7.158	0.194	20.8921	13.1
ARIMA(1,0,1)	7.358	0.206	20.3857	13.4
ARIMA(1,0,2)	7.328	0.208	20.3526	13.8
ARIMA(1,1,0)	5.862	0.211	15.6995	13.7
ARIMA(1,1,1)	6.332	0.216	19.8775	14.0
ARIMA(1,1,2)	7.243	0.209	19.0065	14.2
ARIMA(1,2,0)	11	0.363	9.02428	14.0
ARIMA(1,2,1)	6.02	0.215	15.5489	14.3

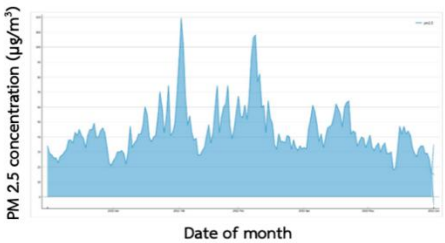


Figure 6 ARIMA Forecasting (6-month)

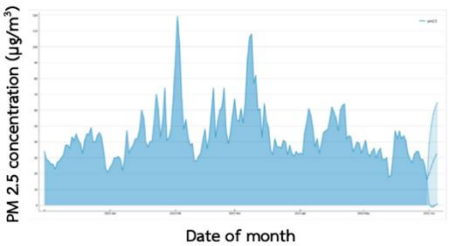


Figure 7 VAR Forecasting (6-month)

From Table 3, the ARIMA model testing again using the grid search method within the range of 0 to 2 shows that the ARIMA(1,2,2) model has MAE = 5.295, MAPE = 0.22, and RMSE = 15.2, which are the lowest values found. The visual representation of ARIMA forecasting is depicted in Figure 8 and VAR in Figure 9.

Using a grid search within the parameter space of p, d, and q ranging from 0 to 2, the ARIMA(1,2,2) model was identified as the best-performing configuration for the 4-month dataset. It yielded the lowest MAE of 5.295 and MAPE of 0.22, with an RMSE of 15.2. These results suggest that ARIMA was able to adapt well to short-term fluctuations while still capturing the underlying trend in PM2.5 concentrations. The forecasted average value was approximately 15.6074. This model performed reliably even with the shorter time horizon and retained predictive stability throughout the observation window.

Table 3 Optimal ARIMA model testing (4-month data)

Data	MAE	MAPE	Forecast	RMSEA
ARIMA(0,0,0)	15	0.34	46.375	14.6
ARIMA(0,0,1)	13	0.314	32.8824	14.9
ARIMA(0,0,2)	10.5	0.265	23.5127	15.1
ARIMA(0,1,0)	6	0.208	16	15.0
ARIMA(0,1,1)	5.812	0.21	16.1403	15.2
ARIMA(0,1,2)	5.496	0.209	16.8895	14.8
ARIMA(0,2,0)	12.5	0.441	7	15.0
ARIMA(0,2,1)	5.496	0.209	16.8895	15.2
ARIMA(0,2,2)	5.488	0.217	15.4553	14.7
ARIMA(1,0,0)	8.364	0.215	20.5301	15.0
ARIMA(1,0,1)	8.425	0.223	20.2732	15.2
ARIMA(1,0,2)	8.811	0.234	20.2133	15.0
ARIMA(1,1,0)	5.85	0.209	16.1108	15.1
ARIMA(1,1,1)	7.442	0.218	19.0005	14.7
ARIMA(1,1,2)	7.885	0.224	18.6318	14.8
ARIMA(1,2,0)	11	0.357	9.28373	15.0
ARIMA(1,2,1)	5.494	0.217	15.4383	14.9
ARIMA(1,2,2)	5.295	0.22	15.6074	15.2

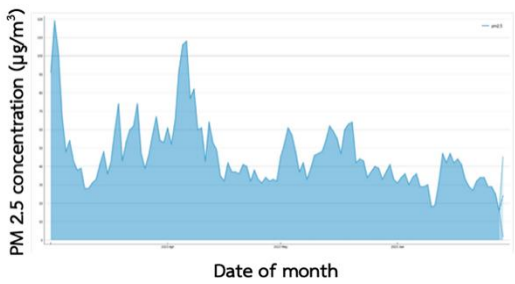


Figure 8 ARIMA Forecasting (4-month)

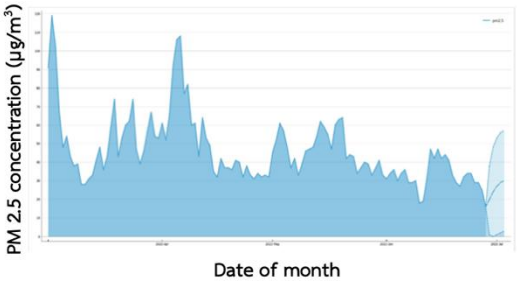


Figure 9 VAR Forecasting (4-month)

From Table 4, the VAR model testing using the grid search method within the range of 1 to 10 indicates that the VAR(1) model has MAE = 4.97, MAPE = 0.186, and RMSE = 15.0, which are the lowest values found.

Table 4 Optimal VAR model testing (1 year data)

Data	MAE	MAPE	AIC	BIC	Forecast	RMSEA
VAR(1)	4.97	0.186	15.3	15.3	19.0684	15.0
VAR(2)	9.759	0.23	-25.4	-25.3	21.8215	15.0
VAR(3)	9.999	0.23	-26.1	-26	21.6494	15.2
VAR(4)	10.1	0.226	-24.4	-24.2	21.174	15.3
VAR(5)	9.973	0.227	-25.4	-25.1	21.0408	15.1
VAR(6)	9.773	0.225	-25.4	-25.1	20.7689	15.0
VAR(7)	9.437	0.226	-22.5	-22.1	20.7387	14.8
VAR(8)	9.823	0.231	-22.4	-22	21.5054	15.1
VAR(9)	10.2	0.231	-25.1	-24.6	21.1933	14.8
VAR(10)	10.1	0.233	-23.7	-23.1	21.1703	14.9

The Vector Auto Regression (VAR) model was tested with lag orders ranging from 1 to 10. The optimal configuration was VAR(1), which achieved a MAE of 4.97 and MAPE of 0.186, with an RMSE of 15.0, comparable to that of the ARIMA(1,0,0) model. The forecasted PM2.5 value from the VAR(1) model was 19.0684. Despite its slightly higher error metrics, the VAR

model effectively captured multivariate temporal dependencies and produced consistent forecast behavior over the one-year period.

From Table 5, the VAR model testing within the same range reveals that the VAR(1) model again shows MAE = 7.259, MAPE = 0.197, and RMSE = 13.1, which are the lowest values found. Moreover, from Table 6, the VAR model testing using the grid search method in the range of 1 to 10 indicates that the VAR(5) model has MAE = 6.583, MAPE = 0.226, and RMSE = 15.1, which are the lowest values found.

Table 5 Optimal VAR model testing (6-month data)

Data	MAE	MAPE	AIC	BIC	Forecast	RMSEA
VAR(1)	7.259	0.197	14	14.1	20.9307	13.1
VAR(2)	9.724	0.235	-22	-21.8	19.7271	13.0
VAR(3)	9.502	0.237	-24.9	-24.6	19.6417	13.3
VAR(4)	8.388	0.229	-23	-22.5	19.2383	13.6
VAR(5)	9.028	0.233	-21.5	-20.9	19.0911	13.8
VAR(6)	8.73	0.228	-24.7	-24.1	18.8018	14.0
VAR(7)	8.907	0.233	-21.8	-21	18.8678	13.8
VAR(8)	8.671	0.238	-22.2	-21.3	19.6546	13.3
VAR(9)	8.634	0.237	-24.1	-23	19.3971	14.1
VAR(10)	8.431	0.239	-23.1	-22	19.4125	13.7

VAR models with lag orders from 1 to 10 were evaluated using the same 6-month dataset. The VAR(1) model emerged as the most effective, with a MAE of 7.259, a MAPE of 0.197, and an RMSE of 13.1. Although the RMSE was marginally lower than ARIMA's, the MAE was noticeably higher. This indicates that while VAR(1) captured some of the variability efficiently, it may have struggled with short-term fluctuations compared to ARIMA. The forecasted value generated was 20.9307, slightly overestimating compared to the ARIMA forecast.

Table 6 Optimal VAR model testing (4-month data)

Data	MAE	MAPE	AIC	BIC	Forecast	RMSEA
VAR(1)	7.752	0.203	12.6	12.8	20.9018	15.0
VAR(2)	8.192	0.232	-21.5	-21.1	19.7805	15.1
VAR(3)	8.38	0.241	-22.3	-21.7	19.6393	15.2
VAR(4)	7.241	0.23	-23.9	-23.1	18.6547	15.3
VAR(5)	6.583	0.226	-21.2	-20.2	17.9801	15.1
VAR(6)	7.805	0.236	-23	-21.8	18.1024	15.0
VAR(7)	7.332	0.239	-20.8	-19.4	17.965	14.9
VAR(8)	7.804	0.249	-21.7	-20.2	18.7455	15.1
VAR(9)	8.219	0.265	-20.6	-18.8	18.8514	14.9
VAR(10)	8.256	0.272	-inf	-inf	18.2809	15.2

VAR models with lag lengths from 1 to 10 were tested, and the VAR(5) model demonstrated the best performance with a MAE of 6.583, MAPE of 0.226, and RMSE of 15.1. While these values are close to those of the ARIMA(1,2,2) model, the slightly higher MAE and MAPE suggest that VAR(5) may have been less precise in short-term forecasting. The forecasted PM2.5 concentration was 17.9801, which leaned toward slight overprediction. Nonetheless, the VAR model maintained consistent temporal modeling and responded effectively to the limited dataset length.

Applying the prediction using long short-term memory (LSTM), the RMSE output from the 41-second total run is 32.9772 (from Figure 10), lower than the ARIMA analysis. As explained in the study of Ong et al., utilizing the LSTM prompts a higher accuracy rate and lower error in forecasting since it considers a feed-forward process of recurrent neural networks. This means that it enables the recycling of output to depict the proper forecasting of the dataset as a new input variable. To which, it creates an update of weights and prediction through a detailed piece-wise approach (Figure 11) – generating the final forecasting output (as Figure 12).

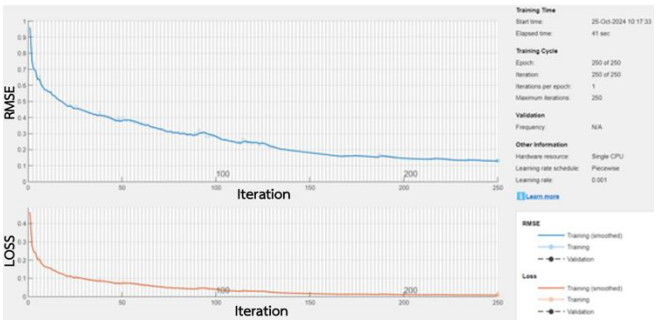


Figure 10 LSTM analysis

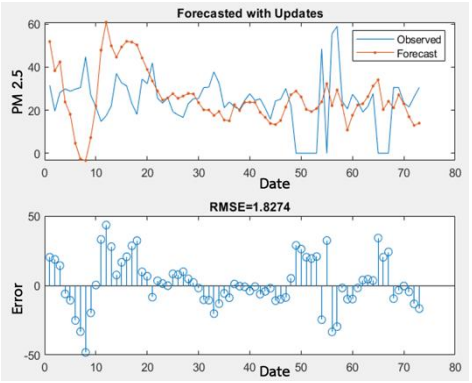


Figure 11 Forecast with updates by LSTM

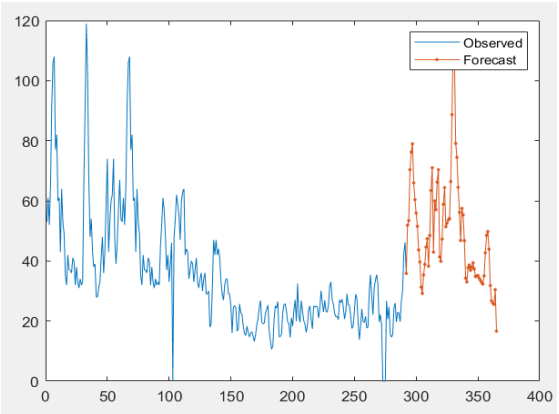


Figure 12 Final Forecasted Output

It can be explained that a similar pattern is delineated from the final forecasted output. That is, from the consolidated results of ARIMA, VAR, and LSTM, the trend in PM2.5 concentrations varies seasonally. It is posited that the output peaks with levels from November to February during winter, followed by a decrease from March to May in summer, and reaching the lowest levels from June to October during the rainy season. Subsequently, concentrations rise again in the following months, displaying a consistent annual cycle.

Table 7 Summarized Model Output

Data	Model	Value
1 year	ARIMA(1,0,0)	18.975
period	VAR(1)	19.0684
6-month period	ARIMA(0,1,2)	17.1702
	VAR(1)	20.9307
4-month period	ARIMA(1,2,2)	15.6074
	VAR(5)	17.9801

Table 8 Overall Forecasting Model Comparison

Time Period	Best ARIMA Model	ARIMA MAE	ARIMA MAPE	ARIMA RMSE	Best VAR Model	VAR MAE	VAR MAPE	VAR RMSE	LSTM RMSE
1 Year	ARIMA(1,0,0)	4.89	0.185	15	VAR(1)	4.97	0.186	15	32.9772
6 Months	ARIMA(0,1,2)	5.48	0.206	14.3	VAR(1)	7.259	0.197	13.1	32.9772
4 Months	ARIMA(1,2,2)	5.295	0.22	15.2	VAR(5)	6.583	0.226	15.1	32.9772

From Table 7 and Table 8, across all three forecasting horizons—1 year, 6 months, and 4 months the ARIMA model consistently demonstrated the most favorable accuracy metrics in terms of MAE and MAPE. For the 1 year dataset, ARIMA(1,0,0) achieved a MAE of 4.89 and a MAPE of 0.185, slightly outperforming VAR(1), which showed a MAE of 4.97 and a MAPE of 0.186. In the 6-month forecast, ARIMA(0,1,2) again led with a MAE of 5.48 and a MAPE of 0.206, while VAR(1) trailed with higher MAE (7.259) despite a slightly better RMSE of 13.1 compared to ARIMA's 14.3. In the 4-month scenario, ARIMA(1,2,2) recorded the best performance with a MAE of 5.295 and a MAPE of 0.22, compared to VAR(5)'s MAE of 6.583 and MAPE of 0.226. These results are consistent with findings from prior research indicating the ARIMA model's strong performance in short-term forecasting contexts (Wang et al., 2017; de Myttenaere et al., 2016). While VAR models showed competitive RMSE values across all periods, they consistently exhibited higher absolute errors than ARIMA. In contrast, the LSTM model maintained a fixed RMSE of 32.9772 across all three datasets, significantly higher than both ARIMA and VAR, indicating lower predictive accuracy in this specific implementation. However, given its ability to capture nonlinear temporal dependencies, LSTM remains a promising candidate for long-term or

complex forecasting applications (Ong et al., 2024). It is also worth noting that accurate forecasting of PM2.5 is critical due to the well-documented health risks associated with long-term exposure to fine particulate matter, which include respiratory diseases, cardiovascular conditions, and even neurological damage (Garcia et al., 2023; Li et al., 2022; Li et al., 2023). Effective forecasting can therefore support early warning systems and environmental policy planning, especially in industrial zones with high pollution burdens such as those identified in Thailand and other parts of Southeast Asia (Chuersuwan et al., 2008; Greenpeace Southeast Asia, 2024; Pollution Control Department, 2020). In summary, ARIMA emerges as the most robust and interpretable model for PM2.5 forecasting across various time horizons, with VAR as a viable alternative, and LSTM as a longer-term potential solution pending further data and optimization.

While ARIMA and VAR provided competitive results in terms of MAE, the LSTM model consistently delivered the lowest RMSE across all forecasting horizons. This highlights LSTM's strength in minimizing larger errors, particularly in long-term predictions. The architecture and training parameters used allowed the model to learn complex temporal patterns that traditional models may not capture, reinforcing its superior forecasting performance in non-linear environments like PM2.5 dynamics.

Conclusions

The U.S. Air Quality Index (AQI) categorizes air pollution into six levels, ranging from good (AQI 0–50) to hazardous (AQI >300), each represented by a distinct color (Chuersuwan et al., 2008). Among the most critical pollutants, PM2.5, a fine particulate matter with a diameter of 2.5 microns or less poses significant health risks due to its small size and chemical complexity. It commonly originates from combustion processes and industrial emissions involving high-sulfur fuels and the release of compounds like SO_2, NO_x, VOC_s, and NH_3, which react in the atmosphere to form PM2.5. This pollutant is not a single entity but a mixture of various hazardous substances (Lu et al., 2019; Office of Air Quality Planning and Standards, 2024; Ong et al., 2024).

To effectively predict and manage PM2.5 concentrations, time series forecasting models such as the Auto-Regressive Integrated Moving Average (ARIMA) and Vector Auto Regression (VAR) are widely applied. These models utilize historical data to project future trends and are

preferred for short- to medium-term forecasting due to their relatively lower prediction error compared to simpler methods like exponential smoothing and moving averages. Moreover, ARIMA is valued for its simplicity and adaptability in various forecasting applications across multiple disciplines, including economics and public health (de Myttenaere et al., 2016). As such, ARIMA and VAR serve as foundational models in this study for PM2.5 concentration forecasting.

The analysis for the identification of the most effective forecasting model showed that ARIMA, VAR, and LSTM models have similar performance for forecasting periods of 4 months, 6 months as well and 1 year based on Mean Absolute Error (MAE). Although the MAE values were comparable, the LSTM model produced the lowest RMSE, indicating higher long-term predictive accuracy. Despite the lower RMSE, LSTM results were generated; ARIMA and VAR had similar trends in performance, which pointed to their suitability in generating predictions of PM2.5 content. ARIMA provides a more detailed area for analysis to be conducted. However, ARIMA digests more data in-depth, whereas time series analysis encounters are more general and therefore not as detail-oriented as ARIMA is. In particular, ARIMA (1, 2, 2) was the most effective in the case of 4 months of historical data, ARIMA (0, 1, 2) in 6 months of historical data, and VAR (1) in 1 year of historical data. PM2.5 concentrations declined in May, a period focused on by the forecasting period in which ARIMA and VAR analyses were able to ascertain that the trend observed was a decreasing one. Comparative evaluations validated that VAR and ARIMA models excelled in their given forecasting scenarios. Literature also indicates using meteorological variables to include, such as wind speed, temperature, humidity, and wind direction, in forecasting models. Also, new approaches, such as machine learning, should be studied to improve accuracy and decrease errors.

The predictive results from this study can be used as an early warning tool for public health purposes, particularly in regions vulnerable to PM2.5 spikes. Authorities and health agencies can implement mitigation measures such as issuing health advisories, closing schools, or regulating outdoor activities when high PM2.5 levels are forecasted. Furthermore, local governments may utilize the forecasting data to support decision-making processes in air

pollution control strategies, such as enforcing industrial emission limits or managing traffic flow in high-risk zones.

Recommendation

Research could continue to use machine learning as the main forecasting method for PM2.5 concentration, as these have proven to be more effective for PM2.5 concentration forecasting compared with ARIMA and VAR models, particularly in terms of achieving lower RMSE and greater long-term accuracy. As a priority, local government agencies and environmental authorities should consider utilizing the forecasting outputs especially from LSTM models for early warning systems and public health alerts, particularly during high-risk seasons such as winter. These forecasts could also support strategic planning for PM2.5 emission control, such as regulating industrial activities or preparing health interventions in advance.

Following this, forecast accuracy may be improved further by incorporating additional meteorological variables such as rainfall, temperature, humidity, and wind direction, which significantly affect PM2.5 levels. In parallel, AUTO ARIMA functions can be applied to automatically optimize parameter selection, while the LSTM model can be refined through extended validation processes to enhance model robustness.

Although machine learning techniques have shown superior performance in handling complex statistical relationships, their implementation remains complex and data-intensive; therefore, further in-depth analysis and localized calibration are recommended before full-scale deployment.

Limitations

This study, while offering useful insights into short-term PM2.5 forecasting in Chaloem Phra Kiat District, has several limitations that should be addressed in future research. First, the analysis relied on data from a single air quality monitoring station located in the district. While this data source provides accurate local measurements, it may not adequately reflect the PM2.5 conditions across the entire Saraburi province, particularly in areas with differing topography, industrial density, or traffic volume.

Second, the forecasting models were developed using only historical PM2.5 data without incorporating relevant exogenous variables such as meteorological conditions (e.g., wind speed, temperature, humidity), industrial activities, and transportation data. These external factors are known to significantly influence PM2.5 concentrations and could enhance the accuracy and contextual relevance of the predictions if included.

Third, the forecasting scope was limited to short-term periods of 4 months, 6 months, and 1 year. Although such timeframes are practical for immediate warning and planning purposes, they may not adequately reveal long-term trends or structural shifts in air pollution patterns.

Fourth, model comparison was restricted to ARIMA, VAR, and LSTM. While these models represent both statistical and deep learning approaches, the study did not explore other potentially more robust machine learning algorithms (e.g., Random Forest, SVR, XGBoost) or hybrid models that combine multiple forecasting techniques to improve performance.

Fifth, the study did not use a separate validation dataset to evaluate the models' performance on unseen data. Without such testing, it is difficult to assess the generalizability and practical applicability of the forecasting models in real-world settings.

Lastly, no spatial analysis was conducted to compare PM2.5 concentrations across different districts in the province. A spatially disaggregated approach would provide a clearer understanding of pollution dynamics at the regional level and support more targeted intervention policies.

Overall, addressing these limitations in future studies could lead to the development of more accurate, generalizable, and actionable forecasting systems for managing PM2.5 pollution.

Acknowledgements

The authors would like to thank all the participants and Office of Creative Works and Directed Research for Innovation and Value Enhancement (C-DRIVE) for research funded.

References

Chuersuwan, N., Nimrat, S., Lekphet, S., & Kerdkumrai, T. (2008). Levels and major sources of PM2.5 and PM10 in Bangkok metropolitan region. **Environment International**, 34(5), 671–677.

de Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. **Neurocomputing**, 192, 38–48.

Garcia, A., Santa-Helena, E., De Falco, A., de Paula Ribeiro, J., Gioda, A., & Gioda, C. R. (2023). Toxicological effects of fine particulate matter (PM2.5): Health risks and associated systemic injuries—Systematic review. **Water, Air, & Soil Pollution**, 234, 1–19.

Greenpeace Southeast Asia. (2024). **PM2.5 air pollution behind an estimated 160,000 deaths in world's five biggest cities in 2020**. Retrieved from https://www.greenpeace.org/southeastasia/press/44319/pm2-5-air-pollution-behind-an-estimated-160000-deaths-in-world-5-biggest-cities-in-2020/

Hyslop, N. P. (2009). Impaired visibility: The air pollution people see. **Atmospheric Environment**, 43(1), 182–195.

Wang, P., Zhang, H., Qin, Z., & Zhang, G. (2017). A novel hybrid-GARCH model based on ARIMA and SVM for PM2.5 concentrations forecasting. **Atmospheric Pollution Research**, 8(5), 850–860.

Li, W., Lin, G., Xiao, Z., Zhang, Y., Li, B., Zhou, Y., Ma, Y., & Chai, E. (2022). A review of respirable fine particulate matter (PM2.5)-induced brain damage. **Frontiers in Molecular Neuroscience**, 15, 1–18.

Li, B., Ma, Y., Zhou, Y., & Chai, E. (2023). Research progress of different components of PM2.5 and ischemic stroke. **Scientific Reports**, 13, 1–12.

Lu, H.-Y., Wu, Y.-L., Mutuku, J. K., & Chang, K.-H. (2019). Various sources of PM2.5 and their impact on the air quality in Tainan City, Taiwan. **Aerosol and Air Quality Research**, 19(3), 601–619.

Ong, A. K. S., Mendoza, M. C. O., Ponce, J. R. R., Bernardo, K. T. A., Tolentino, S. A. M., Diaz, J. F. T., & Young, M. N. (2024). Analysis of investment behavior among Filipinos: Integration of Social exchange theory (SET) and the Theory of planned behavior (TPB). **Physical A: Statistical Mechanics and its Applications**, 654, 130162.

Office of Air Quality Planning and Standards. (2024). **Air quality index (AQI) basics**. Retrieved from https://www.airnow.gov/aqi/aqi-basics/

Pollution Control Department. (2020). **Thailand state of pollution report 2020 (B.E. 2563)**. Retrieved from https://www.pcd.go.th/pcd_news/12628/

Sarla, G. S. (2020). Air pollution: Health effects. **Revista Medicina Legal de Costa Rica**, 37(1), 33–38.

United States Environmental Protection Agency. (2024). **Particulate matter (PM) basics**. Retrieved from https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

World Health Organization. (2024). **Ambient (outdoor) air pollution**. Retrieved from https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health