Accurate Model for Forecasting PM$_{2.5}$ Concentrations in Hat Yai, Songkhla, Thailand: The ARIMA-ANN-REG Hybrid Approach via AAR4PM

Sasira Choojam[1], Jularat Chumnaul[1*], and Korakot Wichitsa-nguan Jetwanna[1,2]

[1] Division of Computational Science, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, Thailand

[2] Air Pollution and Health Effect Research Center, Prince of Songkla University, Hat Yai, Songkhla, Thailand

*Corresponding author: jularat.c@psu.ac.th

Abstract

PM$_{2.5}$ is a significant factor in the troubling air quality that presently affects many countries globally. Therefore, dependable prediction models are needed for the government to enable preparedness for severe PM$_{2.5}$ situations. This study introduces the ARIMA-ANN-REG model to forecast PM$_{2.5}$ concentrations in Hat Yai, Songkhla, Thailand. The efficiency of the proposed model was assessed against the auto-regressive integrated moving average (ARIMA), artificial neural network (ANN), and ARIMA-ANN models. Data used in this study was gathered from the Hat Yai Air Quality Monitoring Station spanning January 1st, 2016, to June 30th, 2022. To develop the model, 70% of the data retrieved between January 1st, 2016, and November 7th, 2020, was utilized for model training, while the remaining 30% was reserved for model testing. In determining the precision of the proposed model, evaluation criteria such as the root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) were employed. The findings showcased the ARIMA-ANN-REG model's superior performance, exhibiting the lowest relative errors among the compared models. In addition, AAR4PM web application was developed for users who want to automatically build the ARIMA, ANN, ARIMA-ANN, and ARIMA-ANN-REG models with their own series data. It can be freely accessed at https://jularatchumnaul.shinyapps.io/AARt4PM/.

Keywords: Time series; Auto-regressive integrated moving average; Artificial neural network; Hybrid model; Forecasting

1. Introduction

Air pollution is one of the most pressing challenges to global health and the environment, ranking as the fourth leading cause of premature death worldwide. Its impact extends as a substantial barrier to the global economy. Particulate matter, specifically PM$_{2.5}$, is a significant factor in causing the troubling air quality. Reports from 2018 revealed that over 64% of over 3,000 cities exceeded the World Health Organization's (WHO) recommended levels of PM$_{2.5}$.

In the same year, Thailand was one of many countries that encountered challenges with PM$_{2.5}$ pollution, surpassing the WHO standard. It ranked 23rd globally and stood third in Southeast Asia, recording a PM$_{2.5}$ concentration of 26.4% (Control Department, 2021). Hat Yai City, in particular, has been grappling with compromised air quality since 2019, attributed to agricultural forest burning and natural forest fires originating from neighboring Indonesia.

Given these circumstances, precise $PM_{2.5}$ concentration forecasting becomes vital for governments to effectively manage pollution levels, ensure compliance with acceptable standards, and issue timely warnings to safeguard residents from hazardous exposure.

In response to the $PM_{2.5}$ issue, extensive research has been conducted in the past five years to develop highly accurate models for forecasting $PM_{2.5}$ concentrations. Linear models such as the auto-regressive (AR), moving average (MA), auto-regressive moving average (ARMA), and auto-regressive integrated moving average (ARIMA) have received extensive exploration for this purpose. For instance, the ARIMA model was employed to analyze and predict $PM_{2.5}$ concentrations in Fuzhou, China, and results indicated seasonal fluctuations in $PM_{2.5}$ concentrations over a two-year period, with higher concentrations observed during cold periods and lower concentrations during warm periods (Zhang *et al.,* 2018). The Holt-Winters additive model was one of the widely used models for forecasting $PM_{2.5}$ concentrations, and it was found that this model exhibited slightly higher prediction accuracy than the seasonal auto-regressive integrated moving average (SARIMA) model (Ao *et al.,* 2019). In a study by Sukkasem (2021) on forecasting $PM_{2.5}$ concentrations in Nakhon Ratchasima, Thailand, the findings suggested that the ARIMA (1, 1, 1) model was suitable for hourly $PM_{2.5}$ concentration forecasts, achieving an accuracy of up to 48 hours. However, for 24-hour average $PM_{2.5}$ forecasting, the ARIMA (0, 1, 3) model outperformed, provide an accuracy of up to 31 days (Sukkasem, 2021). In addition, it was found that the seasonal ARIMA model outperformed the exponential smoothing model in predicting PM2.5 concentrations in Jakarta, exhibiting a slight average percentage difference between predicted values and actual data (Amelia *et al.,* 2022).

Over the past few years, the rise of artificial intelligence (AI) technologies has significantly impacted various domains. AI methods have been extensively studied and applied in many research to address diverse problems, including time series analysis and forecasting. The artificial neural network (ANN) model is one of the models recognized as the highly accurate computational intelligence model for forecasting $PM_{2.5}$ concentrations. In many research, the ANN model was found to be superior to many other models, such as the adaptive neuro-fuzzy inference system (ANFIS) models (Oprea *et al.,* 2017), Support Vector Machine (SVM) models (Masood and Ahmad, 2020), and the multiple linear regression (MLR) models (Bera *et al.,* 2021). Moreover, the ANN model also demonstrated superior forecasting performance compared to the ARIMA model (Liu and You, 2022).

Many hybrid models that combine traditional time series models with AI models were also developed to obtain higher prediction accuracy. The first example of the hybrid model is the ARIMA-ANN proposed by Zhang (2003). This model consists of two components, linear and non-linear, and it has been widely adopted by many researchers (Koutroumanidis *et al.,* 2009; Faruk, 2010; Meth *et al.,* 2010; Khandelwal *et al.,* 2015; Jeerawattanakaset *et al.,* 2017; Naveena *et al.,* 2017; Naveena *et al.,* 2017; Lemuel *et al.,* 2018; Hadwan *et al.,* 2022; Ebrahimi, 2019; Belmahdi *et al.,* 2020). The second example was proposed by Liu and Li (2015). This model combined the results of the ARIMA, ANNs model, and exponential smoothing method (ESM) based on weights derived from the entropy weighting method, and demonstrated excellent performance (Liu and Li, 2015). The third example is the hybrid model proposed by Zhao *et al.* (2022). It is a hybrid ARIMA model based on integrated AIC and improved GS fixed-order methods, along with seasonal decomposition, to forecast $PM_{2.5}$ concentration in Beijing. This model showed significantly improved prediction performance compared to the traditional ARIMA model (Zhao *et al.,* 2022). The last example is the ARIMA-LSTM model which outperformed traditional prediction models in forecasting $PM_{2.5}$ concentration (Xiao *et al.,* 2022). Therefore, this research aimed to propose the ARIMA-ANN-REG model and study its performance compared to the ARIMA, ANN, and ARIMA-ANN models in forecasting daily $PM_{2.5}$ concentration in Hat Yai, Songkhla, Thailand.

Moreover, this study also developed the AAR4PM web application for users who want to automatically build the ARIMA, ANN, ARIMA-ANN, and ARIMA-ANN-REG models with their own series data.

2. Methodology

2.1 Data preparation

The dataset utilized in this study comprised daily $PM_{2.5}$ concentrations, derived as secondary data from the Air Quality and Noise Management Division, spanning from January 1st, 2016, to June 30th, 2022 (Figure 1). The dataset consisted of a total of 2373 observations. However, the data obtained contained certain missing values. Consequently, a data management procedure was conducted to eliminate these missing values, resulting in 1988 observations available for analysis.

To assess the forecasting performance of different models, the daily $PM_{2.5}$ concentrations dataset is divided into two samples of training and testing. According to a rule of thumb for partition data, approximately 70% of the whole dataset is used as training data for model development, and the remaining 30%

is used as test data to evaluate the developed model. The data compositions for the dataset used in this study are given in Table 1.

2.2 Proposed hybrid model for $PM_{2.5}$ prediction

This section presents the proposed hybrid model, Auto-Regressive Integrated Moving Average-Artificial Neural Network-Regression (ARIMA-ANN-REG). The ARIMA-ANN-REG model is initially proposed by Khairalla *et al.* (2017) for forecasting financial time-series data. This model uses the output of the ARIMA and ANN models as independent predictors, and the linear regression analysis is then used to estimate the contribution weight for the two predictors. Therefore, the combination equation can be defined as follows:

$$y_t = w_1 f_t^{ARIMA} + w_2 f_t^{ANN} + \varepsilon_t , \qquad \text{(Eq.1)}$$

where y_t is the observation at time t; f_t^{ARIMA} is the linear component obtained from the ARIMA model; f_t^{ANN} is the non-linear component obtained from the ANN model; w_1, w_2 are the weight values for predictors; ε_t is the random error at time t (Khairalla *et al.*, 2017).
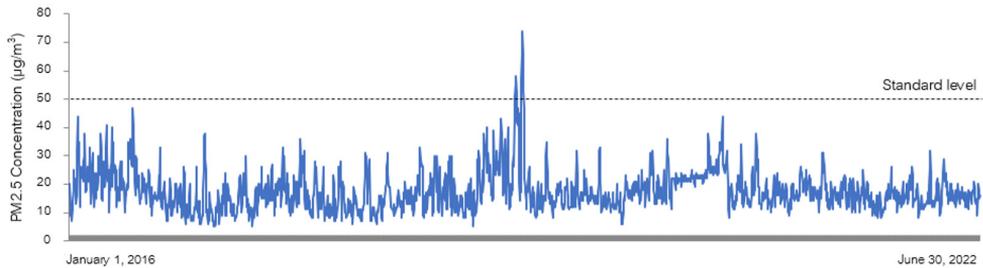

Figure 1. $PM_{2.5}$ concentration (µg/m³) in Hat Yai, Songkhla, Thailand from January 1st, 2016 to June 30th, 2022

Table 1. Sample compositions in the dataset

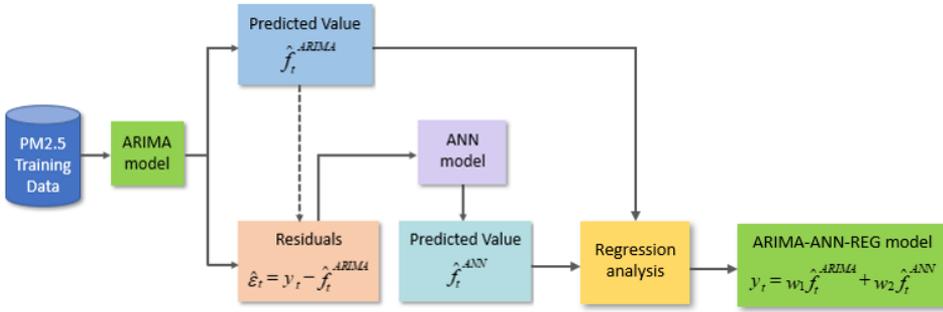| Time series data	*n*	Training set (size)	Test set (size)
Daily $PM_{2.5}$ concentrations	1988	01/01/2016 - 11/07/2020	11/08/2020 - 06/30/2022

Figure 2. The process to obtain the ARIMA-ANN-REG model

Algorithm 1: *Process to obtain the predicted value from the ARIMA-ANN-REG model*

The process to obtain the predicted value from the ARIMA- ANN-REG model, as shown in Figure 2, has 7 steps as follows:

Step 1: Model the linear component for forecasting PM$_{2.5}$ concentrations using the ARIMA model.

Step 2: Determine the predicted PM$_{2.5}$ concentrations at time *t*, denoted as $\hat{f}_t^{ARIMA}$.

Step 3: Determine the residual (e_t) at time *t* from the ARIMA model. The residual can be obtained using the following formula:

$$e_t = y_t - \hat{f}_t^{ARIMA} \qquad (Eq.2)$$

Step 4: Model residuals using the ANN model.

Step 5: Determine the predicted residuals, denoted as $\hat{f}_t^{ANN}$.

Step 6: Use the predicted PM$_{2.5}$ concentrations from Step 2 and the predicted residuals from Step 5 as predictors and determine the regression coefficients as weight values for both predictors, denoted as w_1 and w_2, respectively.

Step 7: Combine the predicted values from Step 2 and Step 5 with the weight in Step 6 to obtain the total forecasting value as shown in the equation below:

$$\hat{y}_t = w_1 \hat{f}_t^{ARIMA} + w_2 \hat{f}_t^{ANN} \qquad (Eq.3)$$

2.3 Models evaluation criteria

Several performance measures have been proposed in the literature to validate forecasting accuracy. It is a common practice to consult various methods to analyze time series data. In this study, we have selected three criteria that possess different properties, which is essential to efficiently understand the forecasting capabilities of the models seen from different points of view. Each of the methods summarized here is a function of the actual daily PM$_{2.5}$ concentration (denoted as y_t) and the forecast value (denoted as $\hat{y}_t$), and *n* refers to the number of observations. The following sections provide a brief overview of each evaluation method used in this study.

2.3.1 Root mean square error (RMSE)

Root mean square error is a measure used to assess the accuracy of a forecast by measuring the difference between the predicted values and the actual observations. A lower RMSE value indicates a higher level of accuracy in the forecast. The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \qquad (Eq.4)$$

2.3.2 Mean absolute error (MAE)

The mean absolute error measures the average error magnitude in a set of forecasts, regardless of direction. It is the mean of the absolute difference test between the forecast and the actual observation, in which the individual differences have the same weight. The MAE formula is given as:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \hat{y}_t\right| \qquad (Eq.5)$$

2.3.3 Mean absolute percentage error (MAPE)

The mean absolute percentage error is scale-independent since it is a percentage error measure of the average absolute error. The forecast is highly accurate if the MAPE is low or close to 0. However, we should be careful when using this criterion because if the time series data have zero values, the MAPE yields undefined results due to division by zero. Moreover, the MAPE is biased towards underestimated values and does not penalize large errors. The MAPE formula is expressed as follows:

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \qquad (Eq.6)$$

2.4 ARIMA-ANN-REG for $PM_{2.5}$ (AAR4Pm) Wep Application

ARIMA-ANN-REG for $PM_{2.5}$ (AAR4Pm) is an interactive and user-friendly web application that can be freely accessed at https://jularatchum naul.shinyapps.io/AAR4PM/ (Figure 3).

In the first step of using AAR4PM, users must prepare their series dataset in .csv format using UTF-8 encoding and upload the prepared data table to the application. Then, users must provide details for model analysis, such as the percentage of training data and the number of periods for forecasting. AAR4PM will automatically build models for the given series data, including ARIMA, ANN, ARIMA-ANN, and ARIMA-ANN-REG, and show the models' performance. The example results of AAR4PM are shown in Figure 4-9.

Figure 3. AAR4PM Web Application

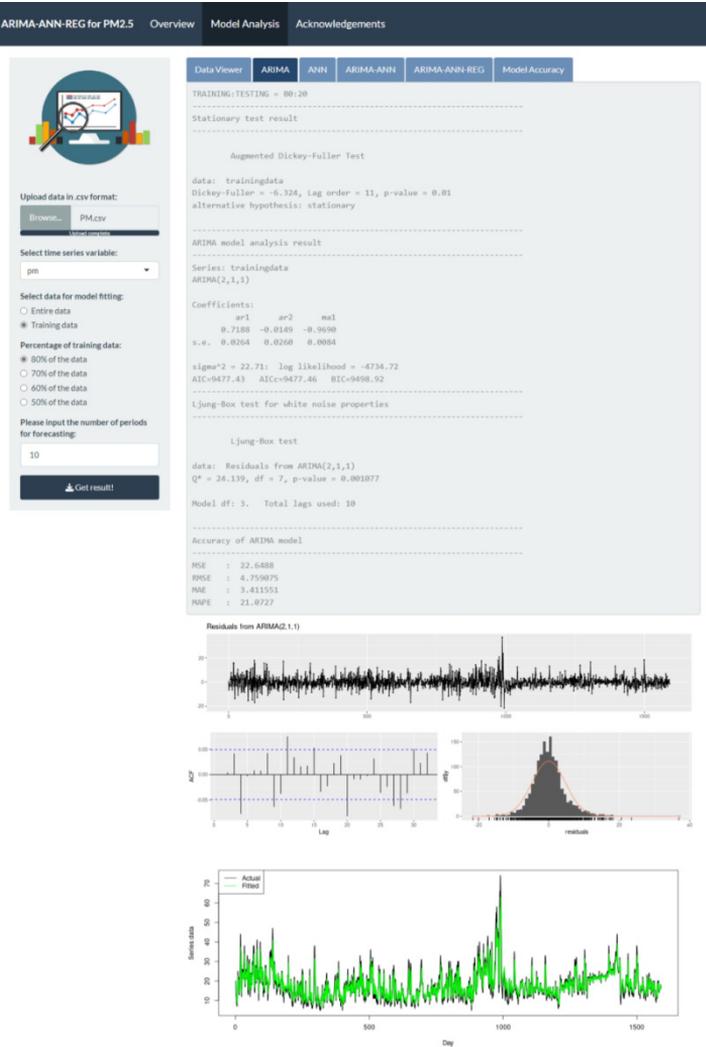**Figure 4.** Results in *Data Viewer* tab

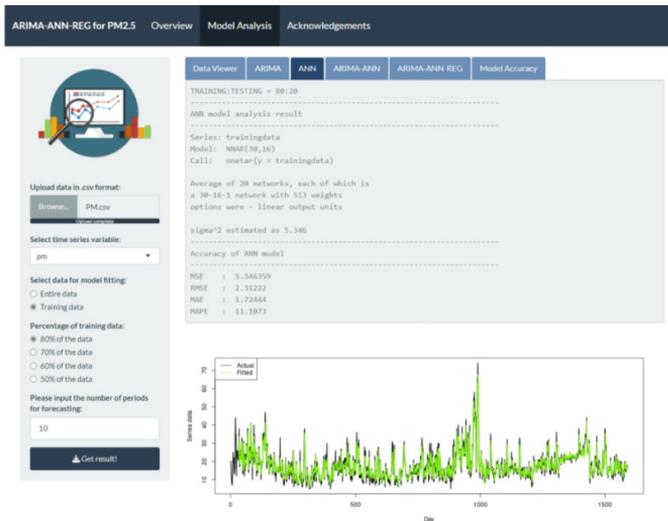**Figure 5.** Results in *ARIMA* tab
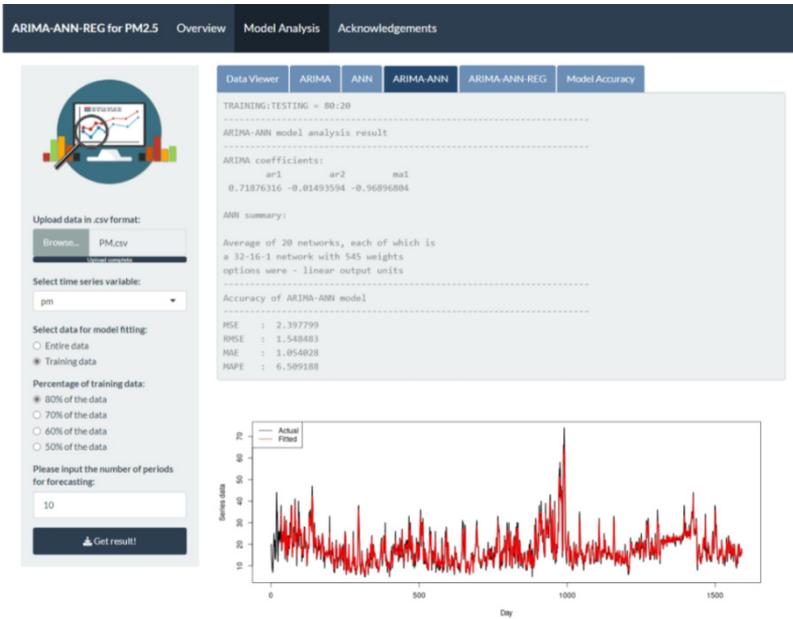

Figure 6. Results in *ANN* tab

Figure 7. Results in *ARIMA-ANN* tab
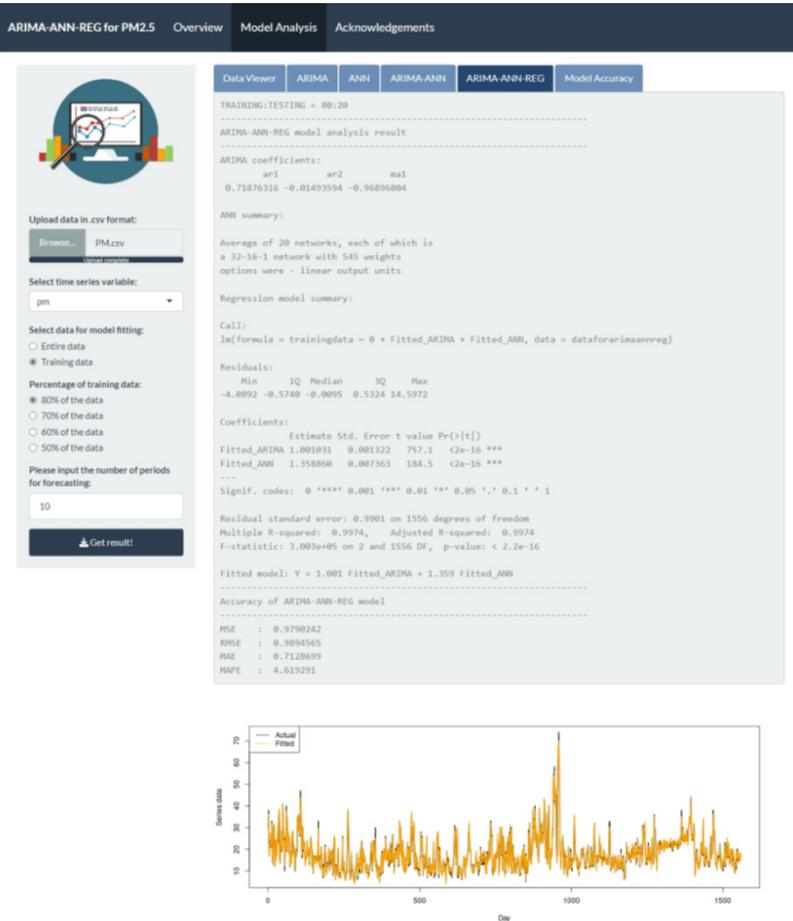

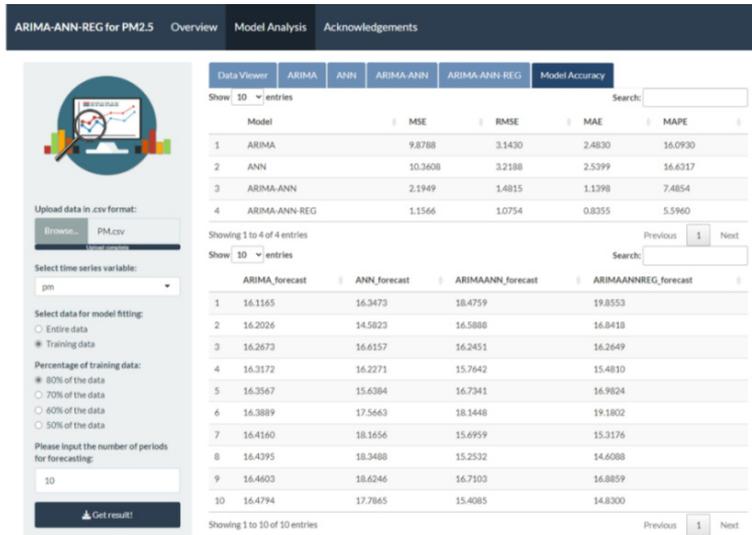**Figure 8.** Results in *ARIMA-ANN-REG* tab

Figure 9. Results in Model Accuracy tab

3. Results and Discussion

3.1 Results

3.1.1 Descriptive statistics for $PM_{2.5}$ concentration

The descriptive statistics of daily $PM_{2.5}$ concentration in Hat Yai, Songkhla, Thailand (from January 1st, 2016, to June 30th, 2022), including the minimum, maximum, mean, standard deviation (SD), and coefficient of variation (CV) are shown in Table 2-4.

According to Table 2, the year 2019 recorded the highest daily $PM_{2.5}$ concentrations in Hat Yai, Songkhla, Thailand, with a mean value of 22.617 µg/m³ (SD = 12.076 µg/m³, and CV = 0.534). The maximum and minimum daily $PM_{2.5}$ concentrations in 2019 were 74 and 8 µg/m³, respectively. Table 3 illustrates that the highest daily $PM_{2.5}$ concentration was in July, with a mean value of 20.762 µg/m³ (SD = 7.066 µg/m³, and CV = 0.622).

Table 4 displays descriptive statistics for the training and testing data of the daily $PM_{2.5}$ concentrations, consisting of 1,391 and 597 observations, respectively. For training data, the mean daily $PM_{2.5}$ concentration was 17.715 µg/m³ (SD = 7.818 µg/m³) with maximum and minimum concentrations of

74 µg/m³ and 5 µg/m³, respectively. On the other hand, the mean daily $PM_{2.5}$ concentration for testing data was 16 µg/m³ (SD = 5.339 µg/m³) with a maximum and minimum of 44 µg/m³ and 8 µg/m³, respectively.

3.1.2 ARIMA model analysis results

In the first step of the ARIMA model analysis, the stationarity of the series data has to be examined. The stationary series is a series in which statistical properties, such as mean, variance, and covariance, remain constant over time or are not a function of time (Montgomery et al., 2015). Here, the Augmented Dickey-Fuller (ADF) test, also known as the Unit Root test, was employed to check such properties. According to the result of the ADF test presented in Table 5, it indicates that the PM2.5 concentration was significantly stationary (p = 0.010).

The next step of the ARIMA model analysis is to find the best ARIMA model. Here, it was found that the most suitable ARIMA model was ARIMA(1, 1, 1) since its AIC value was less than other ARIMA models (AIC = 8359.28) and the p-values for the AR(1) and MA(1) parameters in the ARIMA(1, 1, 1) model were both less than the significance level of 0.05 (Table 6).

Table 2. Descriptive statistics for daily $PM_{2.5}$ concentration ($\mu g/m^3$) by year

Year	Minimum	Maximum	Mean ± SD	CV
2016	6	47	18.824 ± 8.050	0.428
2017	5	36	15.274 ± 5.768	0.378
2018	5	33	15.035 ± 5.731	0.381
2019	8	74	22.617 ± 12.076	0.534
2020	6	44	19.381 ± 5.966	0.308
2021	8	38	16.685 ± 4.574	0.274
2022	8	32	16.129 ± 3.871	0.240

Table 3. Descriptive statistics for daily $PM_{2.5}$ concentration ($\mu g/m^3$) by month

Month	Minimum	Maximum	Mean ± SD	CV
January	5	34	14.900 ± 5.117	0.343
February	7	44	16.601 ± 6.809	0.410
March	6	38	16.411 ± 5.998	0.365
April	5	41	16.753 ± 6.177	0.369
May	8	40	17.017 ± 4.630	0.272
June	8	38	19.188 ± 5.946	0.310
July	9	47	20.762 ± 7.066	0.340
August	8	43	18.220 ± 6.712	0.368
September	7	74	19.790 ± 12.310	0.622
October	5	26	16.773 ± 4.939	0.294
November	7	38	16.122 ± 7.073	0.439
December	6	44	17.392 ± 7.681	0.442

Table 4. Descriptive statistics for training (January 1st, 2016 to November 7th, 2020) and testing (November 8th, 2020 to June 30th, 2022) data

Data	n	Minimum	Maximum	Mean	Median	SD	CV
Training	1391	5	74	17.715	16	7.818	0.441
Testing	597	8	44	17.353	16	5.339	0.308

Table 5. Augmented Dickey-Fuller test

Dickey-Fuller	Lag order	*p*-value
-5.877	11	0.01*

*Significant at the 5% level

Table 6. Parameter estimates and *p*-value of the ARIMA(1,1,1) model for $PM_{2.5}$ concentration

Model	AR(1)	MA(1)	AIC
ARIMA(1,1,1)	0.705	-0.970	8359.28
	0.023*	0.008*	

*Significant at the 5% level

In the diagnostic testing step, results of the one-sample t-test, Bartlett test, and Ljung-Box test in Table 7 show that residuals obtained from the ARIMA(1, 1, 1) had a White Noise feature (*p*-value > 0.05) as depicted in Figure 10. Therefore, the ARIMA(1, 1, 1) model was used to combine with ANN in order to fit the hybrid models proposed in this study.

Table 7. Test the white noise properties of the residual of daily PM$_{2.5}$ concentration in Hat Yai, Songkhla

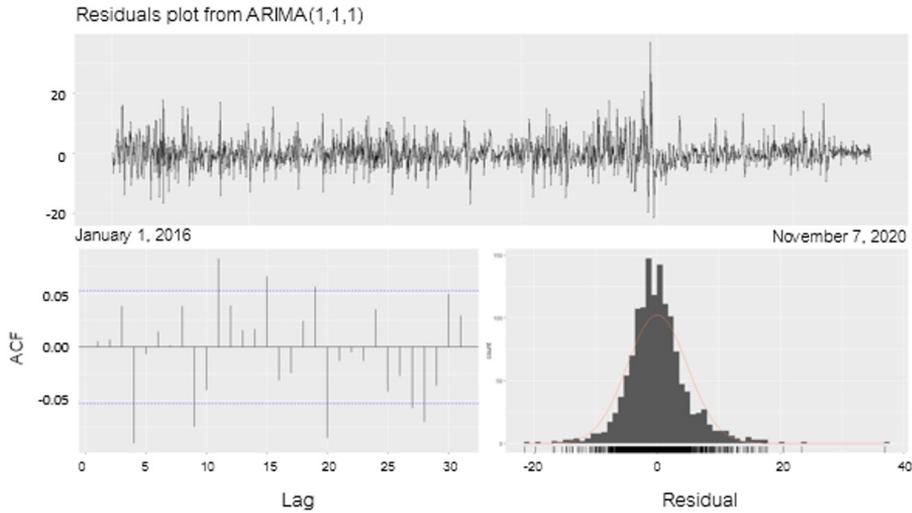| Model	One-sample *t*-test		Bartlett test		Ljung-Box test	
	Test statistic	*p*-value	Test statistic	*p*-value	Test statistic	*p*-value
ARIMA(1,1,1)	0.182	0.856	0.100	0.752	0.038	0.846

Figure 10. White noise properties of the residual

3.1.3 ARIMA-ANN-REG model analysis results

Results of the fitted linear regression model using two predictors derived from the ARIMA(1,1,1) and ANN models are presented in Table 8. Based on the information provided in Table 8, the fitted ARIMA-ANN-REG model can be expressed as follows:

$$\hat{y}_t = 0.988\hat{f}_t^{ARIMA} + 3.243\hat{f}_t^{ANN} \qquad \text{(Eq.7)}$$

The coefficient of determination (adjusted r^2) of the above model was found to be 97.37%, indicating an excellent fitness of the model.

3.1.4 Forecasting analysis and efficiency comparisons

To compare the effectiveness of different models, the ARIMA (1,1,1), and ANN models were initially applied individually to forecast the daily PM2.5 concentrations in Hat Yai, Songkhla, Thailand. Subsequently, the effectiveness of these benchmark models was compared to that of the hybrid models (ARIMA-ANN and ARIMA-ANN-REG) using three evaluation criteria: RMSE, MAE, and MAPE.

Table 9 and Figure 11 summarize the relative errors (RMSE, MAE, and MAPE) of each model based on the training and testing data. According to Lewis's criteria for the MAPE interpretation (Lewis, 1982), it is evident that all the proposed models are good models for forecasting the daily PM$_{2.5}$ concentrations, given that the MAPE values fall within the range of 10-20 (Table 10).

The empirical analysis also confirmed that the relative errors of the ARIMA-ANN-REG were smaller than other models. For the training data, the RMSE, MAE, and MAPE of the ARIMA-ANN-REG model were 1.438, 1.102, and 7.099, respectively. For testing data, the RMSE, MAE, and MAPE of the ARIMA-ANN-REG model were 2.606, 2.033, and 13.558, respectively.

The analysis of relative errors also indicates that the ARIMA-ANN-REG model achieved the highest accuracy in predicting $PM_{2.5}$ concentrations in Hat Yai. Furthermore, empirical analysis confirmed that the ARIMA-ANN and ARIMA-ANN-REG models outperformed. For instance, when evaluating the RMSE for the testing data, both the ARIMA-ANN-REG and ARIMA-ANN models exhibited lower RMSE values (2.606 and 2.757, respectively) compared to the ARIMA model (RMSE = 4.046). This suggests that the combined models had better forecasting accuracy than the ARIMA model alone. Additionally, considering the MAE for the testing data,

both the ARIMA-ANN-REG model (MAE = 2.033) and the ARIMA-ANN model (MAE = 2.130) exhibited more accurate forecasting compared to the ANN model (MAE = 2.664) due to the higher MAE of the ANN model. This further demonstrated the inability of the ARIMA and ANN models to capture all the patterns in the data.

In addition, for the next 7 days forecasting of $PM_{2.5}$, as indicated in Table 11, the ARIMA-ANN-REG model reported the lowest RMSE, MAE, and MAPE of 4.004, 3.444, and 16.595, respectively. Conversely, the ANN model reported the highest values for RMSE, MAE, and MAPE.

Table 8. Results of the multiple linear regression analysis for the ARIMA-ANN-REG model

Model	Predictors	S.E.	t	p-value
ARIMA-ANN-REG	0.988	0.011	90.720	0.000*
	3.243	0.424	7.644	0.000*

*Significant at the 5% level

Table 9. Summary of model accuracy

Model	RMSE		MAE		MAPE	
	Training	Testing	Training	Testing	Training	Testing
ARIMA(1,1,1)	5.020	4.046	3.763	2.884	24.318	16.969
ANN	4.738	3.453	3.431	2.664	21.489	16.566
ARIMA-ANN	1.976	2.757	1.398	2.130	8.742	14.222
ARIMA-ANN-REG	1.438	2.606	1.102	2.033	7.099	13.558

Table 10. Interpretation of typical MAPE values

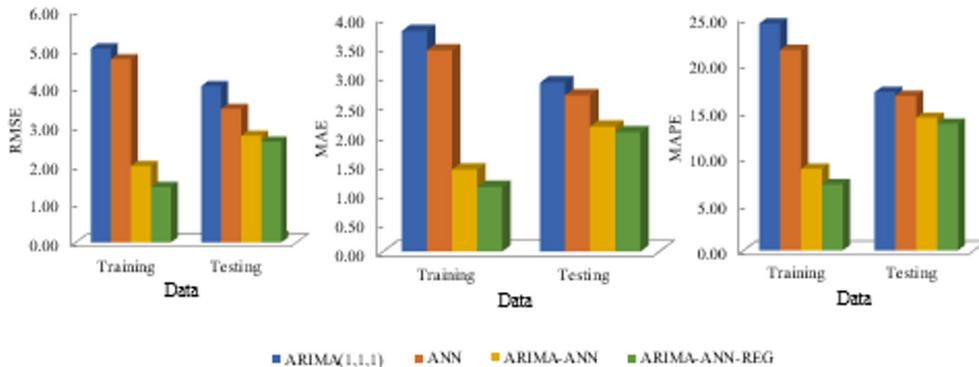| MAPE	Interpretation
Less than 10	Highly accurate forecasting
10 to less than 20	Good forecasting
20 to less than 50	Reasonable forecasting
Greater than 50	Inaccurate forecasting

Figure 11. RMSE, MAPE, and MAE obtained from proposed models

Table 11. Precision of the proposed models in forecasting PM$_{2.5}$ concentrations for the next 7 days (July 1st, 2022 to July 7th, 2022)

Date	PM$_{2.5}$ actual values	PM$_{2.5}$ forecasted values			
		ARIMA	**ANN**	**ARIMA-ANN**	**ARIMA-ANN-REG**
1/7/2022	21	16.117	15.320	15.418	15.055
2/7/2022	21	16.203	13.934	17.558	18.405
3/7/2022	21	16.267	14.459	16.715	17.027
4/7/2022	21	16.317	14.519	15.108	14.445
5/7/2022	17	16.357	13.484	17.042	17.494
6/7/2022	19	16.389	13.736	16.977	17.372
7/7/2022	21	16.416	13.820	17.434	18.083
RMSE		**4.131**	**6.078**	**4.016**	**4.004**
MAE		**3.848**	**5.961**	**3.547**	**3.444**
MAPE		**18.613**	**29.327**	**17.044**	**16.595**

3.2 Discussion

Precise prediction of PM$_{2.5}$ levels holds critical importance for strategic planning in mitigating PM$_{2.5}$ pollution, especially in densely inhabited urban centers like Hat Yai, Thailand. Hat Yai faces heightened vulnerability to elevated PM$_{2.5}$ concentrations, notably intensified during the dry season. Accurate PM$_{2.5}$ forecasting results contribute to several benefits in planning efforts. Firstly, accurate forecasts serve as an early warning system for residents and local authorities. By predicting PM$_{2.5}$ levels in advance, individuals can take preventive measures like staying indoors, using air purifiers, or wearing masks when necessary. Furthermore, local authorities can issue public health advisories and promptly implement pollution control measures (Zhou *et al.,* 2014; He *et al.,* 2023). Secondly, accurate PM$_{2.5}$ forecasts enable healthcare facilities to prepare for potential increases in patients with respiratory and related health issues. Hospitals and clinics can allocate additional resources, such as staff and medication, to efficiently handle the surge in demand during periods of high PM$_{2.5}$ pollution (Yin, 2022). This proactive approach can save lives and alleviate pressure on the healthcare system. Thirdly, accurate forecasting helps transportation authorities anticipate periods of high pollution, primarily caused by vehicular emissions, and implement appropriate traffic management strategies. Measures such as promoting public transportation, encouraging car-pooling, and optimizing traffic flow can effectively reduce emissions, subsequently lowering PM$_{2.5}$ levels in the city. Fourthly, accurate PM$_{2.5}$ forecasting results inform long-term urban planning decisions. City authorities can utilize this data to identify areas consistently experiencing high pollution levels and implement targeted measures to mitigate the problem. For instance, they may consider developing green infrastructure, such as planting trees or creating urban parks, which act as natural filters and improve air quality. Fifthly, accurate PM$_{2.5}$ forecasting provides scientific evidence to support policy implementation aimed at reducing pollution levels. Policymakers gain insights into the severity and impact of PM$_{2.5}$ pollution, leading to the formulation of appropriate regulations and standards to control emissions from industrial activities, power plants, and other pollution sources. Accurate forecasts also facilitate the evaluation of existing policies, allowing for necessary adjustments to enhance their effectiveness (He, *et al.,* 2023). Lastly, accurate PM$_{2.5}$ forecasting results contribute to raising public awareness about the risks associated with high pollution levels. Through public outreach campaigns and educational initiatives, individuals can become informed about the health effects of PM$_{2.5}$ and the preventive measures they can take to safeguard themselves (Minh *et al.,* 2021). This higher awareness can inspire behavioral changes such as reducing personal emissions and fostering support for pollution

control efforts. In summary, accurate $PM_{2.5}$ forecasting results are indispensable for effective planning to combat $PM_{2.5}$ pollution in densely populated cities such as Hat Yai. They enable early warnings, health protection measures, efficient traffic management, informed urban planning decisions, policy implementation based on scientific evidence, and increased public awareness and education. Through these initiatives, Hat Yai can work towards mitigating the impact of $PM_{2.5}$ pollution and improving overall air quality in the city.

4. Conclusion

This study aimed to introduce models for forecasting $PM_{2.5}$ concentrations in Hat Yai, Songkhla, Thailand, and study models' performances. Results of the ARIMA model analysis showed that the most suitable ARIMA model for daily $PM_{2.5}$ concentration in Hat Yai Municipality, Songkhla Province, was ARIMA (1, 1, 1). The RMSE, MAE, and MAPE of training data were found to be 5.020, 3.763, and 24.318, respectively, while the RMSE, MAE, and MAPE of testing data were 4.046, 2.884, and 16.969, respectively.

For the ANN model, the authors used 8 nodes for the input layer, 27 nodes for the hidden layer, and 1 node for the output layer. The maximum number of learning cycles (epochs) was equal to 200, and the learning rate of the network was set at 0.3. Results of the ANN model showed that the RMSE, MAE, and MAPE of the training data were 4.738, 3.431, and 21.489, respectively, and the RMSE, MAE, and MAPE of the testing data were 3.453, 2.664, and 16.566, respectively. Considering the hybrid model (ARIMA-ANN and ARIMA-ANN-REG) proposed in this study, it was found that the ARIMA-ANN-REG model emerged as the most accurate for forecasting daily $PM_{2.5}$ concentration in Hat Yai, whereas the single ARIMA model performed the poorest. Furthermore, the combined models (ARIMA-ANN and ARIMA-ANN-REG) exhibited superior performance compared to the benchmark models (ARIMA, and ANN). These results align with Zhang's (2003) findings, suggesting that combining models (ARIMA-ANN) enhanced forecasting efficiency compared to individual models. Similar findings have been reported by Wang *et al.* (2013), Babu *et al.* (2014), Khandelwal *et al.* (2015), and other researchers (Babu and Reddy, 2015; Elwasify, 2015; Babazadeh, 2017; Lopes *et al.,* 2020; Khan *et al.,* 2021; Sanjeev and Bhardwaj, 2022; Alsuwaylimi, 2023; Siamba *et al.,* 2023), highlighting that hybrid models improved forecasting accuracy over single models. As future research, exploring additional combined models could further enhance forecasting accuracy.

Acknowledgment

References

Alsuwaylimi A. Comparison of ARIMA, ANN and hybrid ARIMA-ANN models for time series forecasting. Information Sciences Letters 2023; 12(2): 1003-1016.

Ameria R. Predicting particulate matter $PM_{2.5}$ using the exponential smoothing and seasonal ARIMA with R studio. IOP Conf. Series: Earth and Environmental Science 2022; 1108(2022): 012079.

Ao X, Yuan H, Zhang D. PM2.5 analysis and prediction based on seasonal time series model, IOP Conf. Series: Earth and Environmental Science 2019; 371: 052006.

Babazadeh R. A hybrid ARIMA-ANN approach for optimum estimation and forecasting of gasoline consumption. RAIRO-Operations Research 2017; 51(3): 719-728.

Babu C, Reddy B. Performance comparison of four new ARIMA-ANN prediction models on internet traffic data. Journal of Telecommunications and Information Technology 2015; 1: 67-75.

Babu C, Reddy B. A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data. Applied Soft Computing 2014; 23: 27–38.

Belmahdi B, Louazazni M, Bouardi A. A hybrid ARIMA–ANN method to forecast daily global solar radiation in three different cities in Morocco. The European Physical Journal Plus 2020; 925(11): 135.

Bera B, Hattacharjee S, Sengupta N, Saha S. $PM_{2.5}$ concentration prediction during COVID-19 lockdown over Kolkata metropolitan city, India using MLR and ANN models. Environmental Challenges (Amst) 2021; 4: 100155.

Ebrahimi A. Time series forecasting of styrene price using a hybrid ARIMA and neural network model. Independent Journal of Management and Production 2019; 10 (3): 915-933.

Elwasify A. A combined model between artificial neural networks and ARIMA models. International Journal of Recent Research in Commerce Economics and Management (IJRRCEM) 2015; 2(2): 134-140.

Faruk D. A hybrid neural network and ARIMA model for water quality time series prediction. Engineering Applications of Artificial Intelligence 2010; 23: 586–594.

Hadwan M, Al-Maqaleh B, Al-Badani FN, Khan RU, Al-Hagery MA. A hybrid neural network and Box-Jenkins models for time series forecasting. Computers, Materials and Continua 2021; 07(03): 4829-4845.

He J, Li X, Chen Z, Mai W, Zhang C, Wan X, Wang X, Huang M. A hybrid CLSTM-GPR model for forecasting particulate matter (PM2.5). Atmospheric Pollution Research 2023; 14(8): 101832.

Jeerawattanakaset E, Ning X, Al-Jallad N. Hybrid forecasting scheme for financial time-series data using neural network and statistical methods. International Journal of Advanced Computer Science and Applications 2017; 8(9): 319-327.

Khairalla M, Xu-Ning, Al-Jallad NT. Hybrid forecasting scheme for financial time-series data using neural network and statistical methods. IJACSA 2017; 8: 319-327.

Khan F, Urooj A, Muhammadullah S. An ARIMA-ANN hybrid model for monthly gold price forecasting: Empirical evidence from Pakistan. Pakistan Economic Review 2021; 1(1): 61–75.

Khandelwal I, Adhikari R, Verma G. Time series forecasting using hybrid ARIMA and ANN models based on DWT decomposition. Procedia Computer Science 2015; 48: 173-179.

Koutroumanidis T, IoANNou K, Arabatzis G. Predicting fuel wood prices in Greece with the use of ARIMA models artificial neural networks and a hybrid ARIMA-ANN model. Energy Policy 2009; 37: 3627-3634.

Kulvanich N. Improving prediction accuracy of time series data using ARIMA-ANN hybrid model and regression analysis. Thai Science and Technology Journal 2022; 28(12): 2101–2112.

Lewis CD. Industrial and business forecasting methods. Butterworths Publishing, London, UK. 1982.

Liu D, Li L. Application study of comprehensive forecasting model based on entropy weighting method on trend of PM2.5 concentration in Guangzhou, China. International Journal of Environmental Research and Public Health 2015; 12(6): 7085-7099.

Liu T, You S. Analysis and forecast of Beijing's air quality index based on ARIMA model and Neural Network model. Atmosphere 2022; 13(4): 512.

Lopes RLF, Fraiha SGC, Gomes HS, Lima VD, Cavalcante GPS. Application of hybrid ARIMA and artificial neural network modeling for electromagnetic propagation: An alternative to the least squares method and ITU recommendation P.1546-5 for Amazon urbanized cities. Applied Soft Computing 2020; 8494185.

Masood A, Ahmad K. A model for particulate matter ($PM_{2.5}$) prediction for Delhi based on machine learning approaches. Procedia Computer Science 2020; 167: 2101-2110.

Meth N, Saxena V, Pardasani K. A comparison between hybrid approaches of ANN and ARIMA for Indian stock trend forecasting. Business Intelligence Journal 2010; 3: 23-43.

Minh VTT, Tin TT, Hien TT. $PM_{2.5}$ forecast system by using machine learning and WRF model, a case study: Ho Chi Minh City, Vietnam. Aerosol and Air Quality Research 2021; 21(12): 210108.

Montgomery D, Jennings C, Kulahci M. Introduction to time series analysis and forecasting. John Wiley and Sons, New Jersey, USA. 2015.

Naveena K, Singh S, Rathod S, Singh A. Hybrid ARIMA-ANN modelling for forecasting the price of robusta coffee in India. International Journal of Current Microbiology and Applied Sciences 2017; 6(7): 1721-1726.

Oprea M, Mihalache SF, Popescu M. Computational intelligence-based $PM_{2.5}$ air pollution forecasting. International Journal of Computers Communications and Control 2017; 12(3): 365-380.

Sanjeev, Bhardwaj N. Time series prediction using hybrid ARIMA-ANN models for sugarcane. International Journal of Plant and Soil Science 2022; 34(23): 772-782.

Siamba S, Otieno A, Koech J. Application of ARIMA, and hybrid ARIMA models in predicting and forecasting tuberculosis incidences among children in Homa Bay and Turkana counties, Kenya. PLOS Digital Health 2023; 2(2): e0000084.

Sukkasem I. The amount of particulate matter that have a diameter of less than 2.5 micrometers ($PM_{2.5}$) forecasted in Nakhon Ratchasima Province. The Office of Disease Prevention and Control 9th Nakhon Ratchasima Journal 2021; 27(1): 16-25.

Velasco LCP, Polestico DLL, Macasieb GPO, Reyes MBV, Vasquez FB. Load forecasting using autoregressive integrated moving average and artificial neural network. International Journal of Advanced Computer Science and Applications 2018; 9(7): 23-29.

Wang L, Zou H, Su J, Li L, Chaudhry S. An ARIMA-ANN hybrid model for time series forecasting. Systems Research and Behavioral Science 2013; 30(3): 244-259.

Xiao J, Wang Q, Cui J, Yu J. Multi-feature $PM_{2.5}$ prediction with ARIMA-LSTM. 2022 International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM), Xiamen, China, 2022; 229-233.

Yin S. Decadal changes in $PM_{2.5}$-related health impacts in China from 1990 to 2019 and implications for current and future emission controls. Science of the Total Environment 2022; 834: 155334.

Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 2003; 50: 159-175.

Zhang L, Lin J, Qiu R, Hu X, Zhang H, Chen Q, Tan H, Lin D, Wang J. Trend analysis and forecast of $PM_{2.5}$ in Fuzhou, China using the ARIMA model. Ecological Indicators 2018; 95(Part 1): 702-710.

Zhao L, Li Z, Qu L. Forecasting of Beijing $PM_{2.5}$ with a hybrid ARIMA model based on integrated AIC and improved GS fixed-order methods and seasonal decomposition. Heliyon 2022; 8(12): e12239.

Zhou Q, Jiang H, Wang J, Zhou J. A hybrid model for $PM_{2.5}$ forecasting based on ensemble empirical mode decomposition and a general regression neural network. Science of the Total Environment 2014; 496: 264-274.