International Journal of
Statistics and Applied Mathematics

**ISSN:** 2456-1452
Maths 2023; 8(3): 258-264
© 2023 Stats & Maths
https://www.mathsjournal.com
Received: 09-02-2023
Accepted: 11-03-2023

**Jinrun Zhong**
College of Mathematics and Statistics, Guangdong University of Foreign Studies, Guangzhou, China

**Bo Cheng**
College of Mathematics and Statistics, Guangdong University of Foreign Studies, Guangzhou, China

**Corresponding Author:**
**Bo Cheng**
College of Mathematics and Statistics, Guangdong University of Foreign Studies, Guangzhou, China

# Predicting concentrations of atmospheric particle matters in Guangzhou by time series models

**Jinrun Zhong and Bo Cheng**

**DOI:** https://doi.org/10.22271/maths.2023.v8.i3c.1042

**Abstract**
Particulate matter is one of the major air pollutants closely related to human health. In order to predict atmospheric particulate matter concentrations effectively and accurately, this paper utilized ARIMA model, Holt-Winters model, STL-Holt model and STL-ARIMA model to carry out prediction experiments based on hourly PM2.5 and PM10 concentration historical data in Guangzhou city. The results showed that the four models were effective in predicting hourly PM2.5 and PM10 concentrations. The RMSE, MAE, MAPE, and $R^2$ metrics were used to evaluate the prediction accuracy of the models. It was found that the Holt-Winters model performed best among the four models. This study may provide guides for the environmental authorities in forecasting atmospheric particulate matter concentrations.

**Keywords:** Particulate matter prediction, ARIMA, holt-winters, STL

## 1. Introduction

With rapid economic and social development, the issue of air pollution has become a major concern for us. Air pollution has a constant impact on people's health and travelling. Long-term exposure to atmospheric particulate matter in concentrations exceeding normal standards can lead to cardiovascular disease, ischaemic heart disease, and a range of diseases related to the lung and respiratory tract (Kim *et al*., 2017; Miller *et al*., 2007) [1, 2]. Especially after the COVID-19 pandemic, people's hearts and lungs may not function as well as before, putting them at a much higher risk of developing a disease. An effective air quality forecast will not only reduce the adverse health effects of air pollution for individuals but also enable government departments to carry out management work such as reducing road emissions and increasing forest cover (Simiyu *et al*., 2021) [3], thus providing significant economic and ecological benefits to society.

Particulate matter (PM) is one of the major air pollutants released into the atmosphere as a result of human activities. Models for predicting atmospheric particulate matters (PM2.5 and PM10) concentrations mainly include statistical models, chemical transport models and machine learning models. Among the statistical models, time series models are the conventional but effective prediction methods. Bhatti *et al*. (2021) [4] used a SARIMA model to fit monthly PM2.5 concentrations in Lahore city of Pakistan from 2014 to 2020 and predicted the data for the next 12 months. The results showed that 75% of the predictions had a percentage error of less than 7%. The Brazilian scholar Ventura *et al*. (2019) [5] proposed using the Holt-Winters model to predict air quality and got better results in the simulation prediction of PM2.5 in the industrial area of Rio de Janeiro. Wongrin *et al*. (2023) [6] used statistical methods and deep learning techniques to predict daily average PM2.5 concentrations in northern Thailand and showed that ARIMA and ETS models performed better than deep learning methods at most stations. Aladağ (2021) [7] combined wavelet transform and ARIMA models to construct a WT-ARIMA model to predict monthly PM10 concentrations in Erzurum, Turkey, and indicated that the hybrid model gave better predictions than the traditional ARIMA model.

However, little work has been done to apply the Holt-Winters model and the STL decomposition method to predict atmospheric particulate matter concentrations in Chinese cities. In this paper, the ARIMA model, Holt-Winters model and STL decomposition method will be applied to predict and analyse the hourly PM2.5 and PM10 concentrations in Guangzhou one of the largest cities in China, which may provide references for the prevention and control of atmospheric particulate matters pollution.

## 2. Materials and Methods

### 2.1 Data acquisition and processing

This paper selects Guangzhou's hourly particulate matter concentration data in January 2022 for the study, including PM2.5 and PM10 concentration data. Guangzhou is a large city in southern China with an area of 7434.40 square kilometres and a resident population of 18,734,100 at the end of 2022. The particulate matter concentration data are available from the China Air Quality Historical Data website (http://beijingair.sinaapp.com). There are outliers and missing values in the data that need to be dealt with. For the outliers, this paper takes the way of eliminating and making them as missing values. For the missing values, this paper takes the linear interpolation method to fill in, making the time series data more complete. All the analyses are based on the statistical software R in version 4.2.2.

### 2.2 Research Methodology

**ARIMA Model**

The ARIMA model is a time series forecasting method proposed by the American statistician Box and the British statistician Jenkins in 1976. It is a generalized form of the Auto Regressive Moving Average (ARMA). If the time series $\{y_t\}$ is stationary and satisfies the equation (Cryer and Chan, 2008) [8]:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

$\{y_t\}$ is called a autoregressive moving average process of orders p and q; we abbreviate the name to $ARMA(p,q)$, where $\phi_1,\cdots,\phi_p$ are Auto Regressive coefficients, and $\theta_1,\cdots,\theta_q$ are the Moving Average coefficients. A non-stationary time series $\{y_t\}$ is said to be an $ARIMA(p,d,q)$ process if the DTH difference $W_t = \nabla^d Y_t$ is a stationary $ARMA(p,q)$ process.

The modelling process for the ARIMA model is as follows: (1) Data testing: Before building the model, the data need to be tested for stationarity. Series that are not stationary need to be differenced first so that they meet the stationarity requirements for ARIMA modelling. A white noise test is also required to ensure that the series is not completely random. Only if the series is not a white noise process does it make sense to build a model for analysis and prediction. (2) Model identification: The parameters in the $ARIMA(p,d,q)$ model can be investigated by autocorrelation and partial autocorrelation graphs of the stationary time series and the Akaike information criteria (AIC). The model with the smaller AIC value is preferred as the best form of the model. (3) Model diagnosis: After building the ARIMA model, the residual series fitted to the model need to be tested for white noise. Models that pass the test can be used for forecasting, while those that do not need to be rebuilt.

**Holt-Winters Model**

Holt-Winters model is a three-parameter exponential smoothing method that was proposed by Winters to capture seasonality in 1960. The Holt-Winters method consists of a forecasting equation and three smoothing equations - one for the level $l_t$, one for the trend $b_t$, and one for the seasonal component $s_t$. According to the different seasonal components, there are additive and multiplicative models for this method. When seasonal changes in the time series remain roughly constant, the additive model is usually preferred, while when seasonal changes vary proportionally with the level of the time series, the multiplicative model is usually preferred (Hyndman and Athanasopoulos, 2018) [9]. Assuming that $\hat{y}_{t+h|t}$ is the predicted value of the series $\{y_n\}$ at time $h$ and $m$ is the frequency of seasonality, then the component form for the additive method is:

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)}$$

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1})$$
$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$
$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$$

The component form for the multiplicative method is:

$$\hat{y}_{t+h|t} = (l_t + hb_t)s_{t+h-m(k+1)}$$

$$l_t = \alpha(\frac{y_t}{s_{t-m}}) + (1-\alpha)(l_{t-1} + b_{t-1})$$
$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$
$$s_t = \gamma(\frac{y_t}{l_{t-1} + b_{t-1}}) + (1-\gamma)s_{t-m}$$

Where $k$ is the integer part of $(h-1)/m$ which ensures that the estimate of the seasonality index used for prediction is from the last period of the sample. The usual restriction on the model smoothing parameters is $\alpha, \beta, \gamma \in [0,1]$.

### Forecasting with STL decomposition

The STL (Seasonal and Trend decomposition using Loess) is a versatile and robust method for time series decomposition developed by Cleveland *et al*. (1990) [10]. We denote $y_t$ as the sample data, $T_t$ as the trend component, $S_t$ as the seasonal component and $R_t$ as the remainder, then the additive decomposition of the time series can be written as

$$y_t = T_t + S_t + R_t, t = 1, \cdots, N.$$

Assuming a decomposed time series is written as $y_t = S_t + A_t$, where $A_t = T_t + R_t$ is the seasonally adjusted component. To forecast a decomposed time series, the seasonal component $S_t$ and the seasonally adjusted component $A_t$ need to be predicted separately (Hyndman and Athanasopoulos, 2018) [9]. The seasonal component is usually assumed to be constant or to vary very slowly, so it is predicted by simply extracting the last period of the estimated component. For the seasonally adjusted component, any non-seasonal forecasting method can be utilized, such as the random walk with drift model, the ARIMA method and the Holt method.

### Model Assessment

In this paper, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Correlation Coefficient $R^2$ are selected to evaluate the prediction accuracy of the models. They are defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|$$
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$
$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\%$$
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i - y_i)^2}$$

where n is the number of samples. The MAE, RMSE, MAPE and $R^2$ are used to measure the errors between predicted values ($\hat{y}_i$) and real values ($y_i$).

## 3. Results and Discussions

### Descriptive Statistics

The results of the descriptive statistical analysis of the observed particulate matter concentration data are shown in Table 1. The sample size for the hourly PM2.5 and PM10 concentrations data used in this study was 504. The mean and standard deviation of PM2.5 concentrations were 37.56 and 16.00 μg/m$^3$, with Skewness and Kurtosis values calculated as 0.20 and -0.39; the mean and standard deviation of PM10 concentrations were 63.79 and 29.83 μg/m$^3$, with skewness and kurtosis values calculated as 0.40 and -0.53. According to the Jarque-Bera normality test, the results showed that the p-values of the two time series were smaller than 0.05, rejecting the normal distribution of the observed data.

**Table 1:** Descriptive Statistics Summary

| Variables | PM2.5 | PM10 |
|---|---|---|
| Observations | 504 | 504 |
| Range | [4,78] | [6, 143] |
| Median | 37 | 60 |
| Mean | 37.56 | 63.79 |
| Std. Dev | 16.00 | 29.83 |
| Skewness | 0.20 | 0.40 |
| Kurtosis | -0.39 | -0.53 |
| Jarque-Bera.p | 0.043 | 0.000 |

## Modelling process

### ARIMA Model

**1. Data Testing:** To check the stability, a unit root ADF (Augmented Dickey-Fuller) test was conducted on the hourly PM2.5 concentration and PM10 concentration data for the first 20 days of January 2022 in Guangzhou. The results in Table 2 showed that both series of PM2.5 concentration (series I) and PM10 concentration (series II) were non-stationary series, but the p-values of the ADF test for the two series after the first-order difference were smaller than 0.01, which meant the new series diff_PM2.5 and diff_PM10 were stationary series. The Box-Ljung test showed that the p-values of both series are smaller than 0.01, rejecting the series as completely random. It suggested that the two series are not white noise and an ARIMA model can be built.

**Table 2:** Unit root test (Augmented DF)

| Variables | Test statistic | p-Value |
|---|---|---|
| PM2.5 | -2.53 | 0.35 |
| diff_PM2.5 | -7.88 | < 0.01 |
| PM10 | -2.83 | 0.23 |
| diff_PM10 | -8.08 | < 0.01 |

**2. Model Selection and Diagnosis:** First, the parameters of the ARIMA model were automatically selected in R software using the autoarima command to identify the model for series I as SARIMA (2, 1, 1) (0, 0, 2)$_{24}$ with an AIC value of 1975.94, and the model for series II as SARIMA (5, 1, 0) (0, 0, 2)$_{24}$ with an AIC value of 2510.08.

Later, the parameters of the ARIMA model were adjusted by ACF and PACF graphs. Figure 1 showed the ACF and PACF graphs for the stationary series after a first difference. The ACF graph for series I displays significance at lags 1, 2, 23 and 24, and the PACF graph displays significance at lags 1, 2 and 24. Hence the model parameters for series I are adjusted to SARIMA (2, 1, 2) (1, 0, 2)$_{24}$ with an AIC value of 1962.81, which is smaller than the result before the adjustment. The ACF graph for series II displays significance at lags 1 and 2. The model parameters were therefore adjusted to SARIMA (5, 1, 2) (0, 0, 2)$_{24}$ with an AIC value of 2509.27 and a BIC value of 2550.98. The AIC value was similar to the model before the adjustment, but the BIC value was larger than before. Therefore, as shown in Table 3, the final models for the two series were identified as SARIMA (2, 1, 2) (1, 0, 2)$_{24}$ and SARIMA (5, 1, 0) (0, 0, 2)$_{24}$ according to the AIC and BIC criterion. The Ljung-Box test for the residual series of the two models showed that the p-values were larger than 0.05, suggesting that the residuals are white noise series.

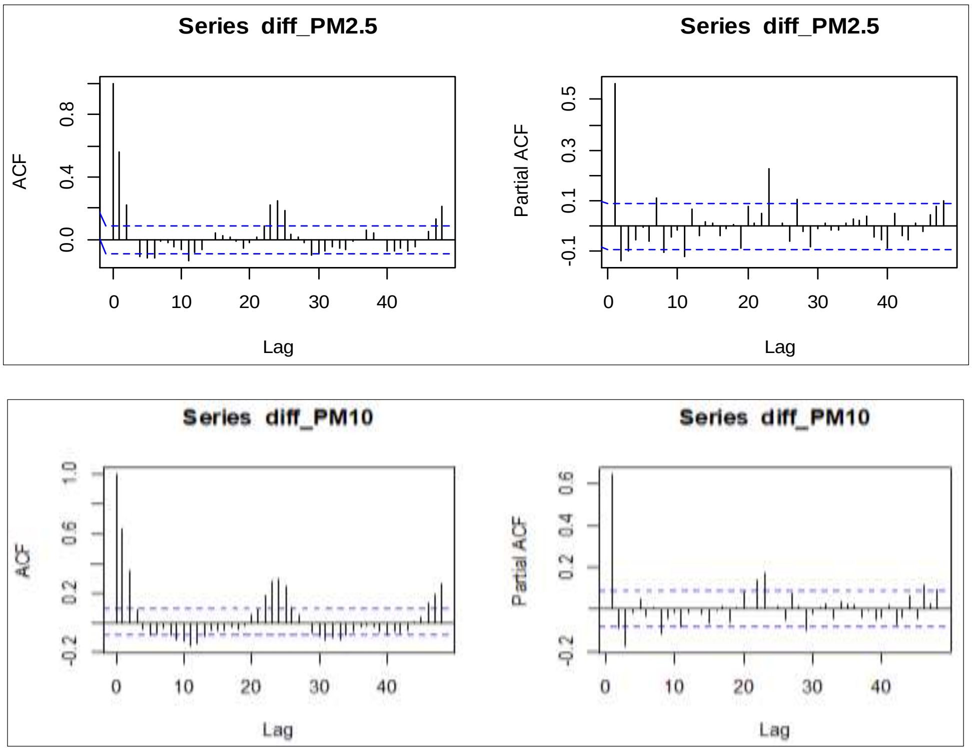

**Fig 1:** ACF and PACF graphs for series after a first difference

**Table 3:** ARIMA Models for Atmospheric Particle Matters

| Particle Matters | ARIMA Model | AIC | BIC |
|---|---|---|---|
| PM2.5 | SARIMA(2,1,1)(0,0,2)$_{24}$ | 1975.94 | 2000.97 |
| | SARIMA(2,1,2)(1,0,2)$_{24}$ | 1962.81 | 1996.19 |
| PM10 | SARIMA(5,1,0)(0,0,2)$_{24}$ | 2510.08 | 2543.46 |
| | SARIMA(5,1,2)(0,0,2)$_{24}$ | 2509.27 | 2550.98 |

### Holt-Winters Model

The observations of PM2.5 hourly concentrations and PM10 concentrations were fitted with Holt-Winters additive and multiplicative models, and the smoothing parameters of the models were calculated with R software. As shown in Table 4, the AIC values of the additive model were smaller than those of the multiplicative model in fitting the two particle concentrations. Therefore, the Holt-Winters additive model was finally selected.

**Table 4:** Comparison of the AIC values and parameters

| Models | PM2.5 | | | | PM10 | | | |
|---|---|---|---|---|---|---|---|---|
| | AIC | $\alpha$ | $\beta$ | $\gamma$ | AIC | $\alpha$ | $\beta$ | $\gamma$ |
| additive | 3808 | 0.955 | 0.03 | 0.045 | 4391 | 0.962 | 0.002 | 0.037 |
| multiplicative | 3963 | 0.925 | 0.003 | 0.074 | 4574 | 0.964 | 0.034 | 0.036 |

### Forecasting with STL decomposition

The hourly PM2.5 concentration and PM10 concentration observations were decomposed using the STL method. As can be seen in Figure 2, the seasonal component of the decomposition showed that the series fluctuates over time, indicating seasonality in both series. In this paper, we forecast the non-seasonal components of the two series using the Holt method and the ARIMA method, where the ARIMA models are in the form of ARIMA (2, 1, 2) and ARIMA (5, 1, 2).

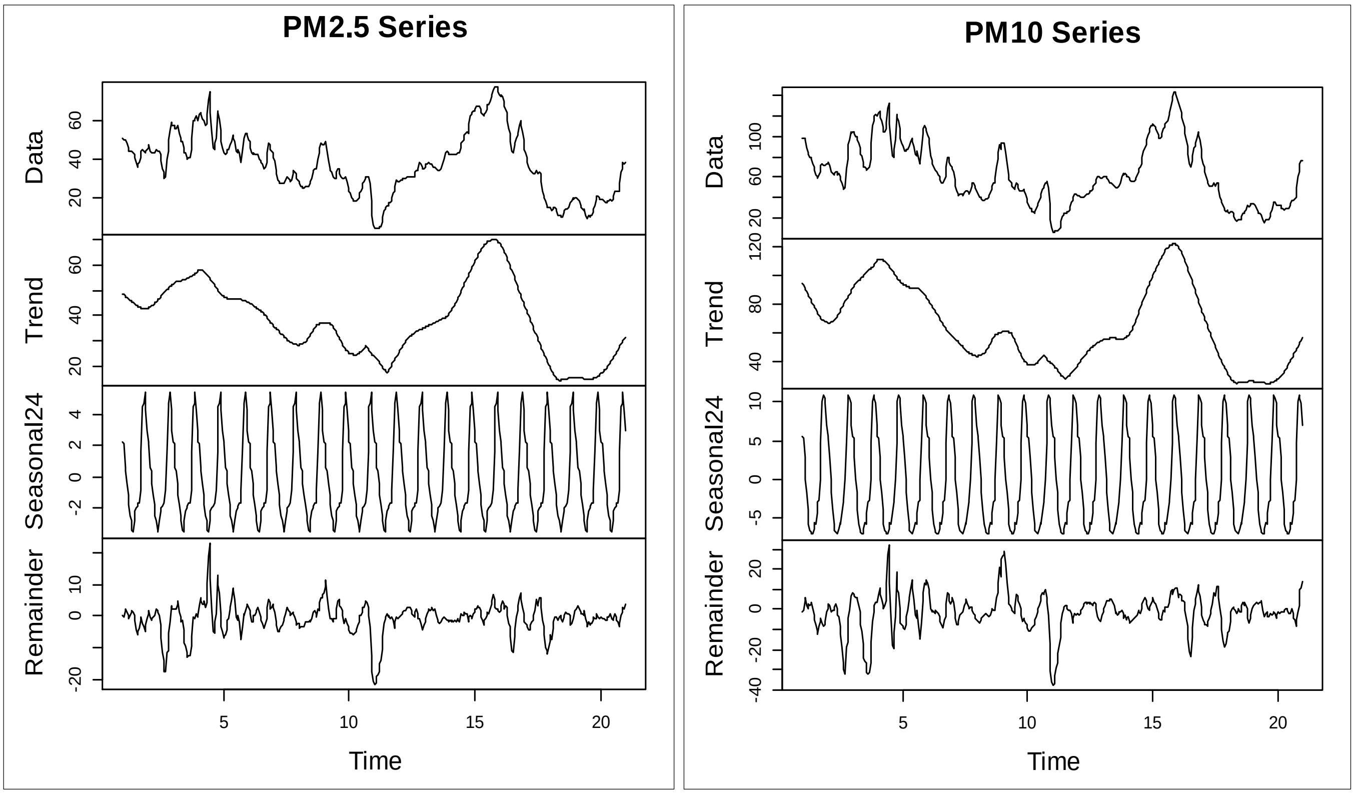

**Fig 2:** Time series decomposition of PM2.5 and PM10

### Model Prediction Results

The established SARIMA model, Holt-Winters model and STL decomposition method were used to fit the hourly PM2.5 and PM10 concentration data for the first 20 days of January 2022 in Guangzhou city and to predict the concentrations for the next 24 hours. The actual values were compared with the predicted values of the models in Figure 3 and Figure 4. It can be seen from the figures that there is a consistent upward trend in the predicted values and the true values, which indicates that the models have a favorable predictive effect. It is observed that the Holt-Winters method performs best in the prediction.

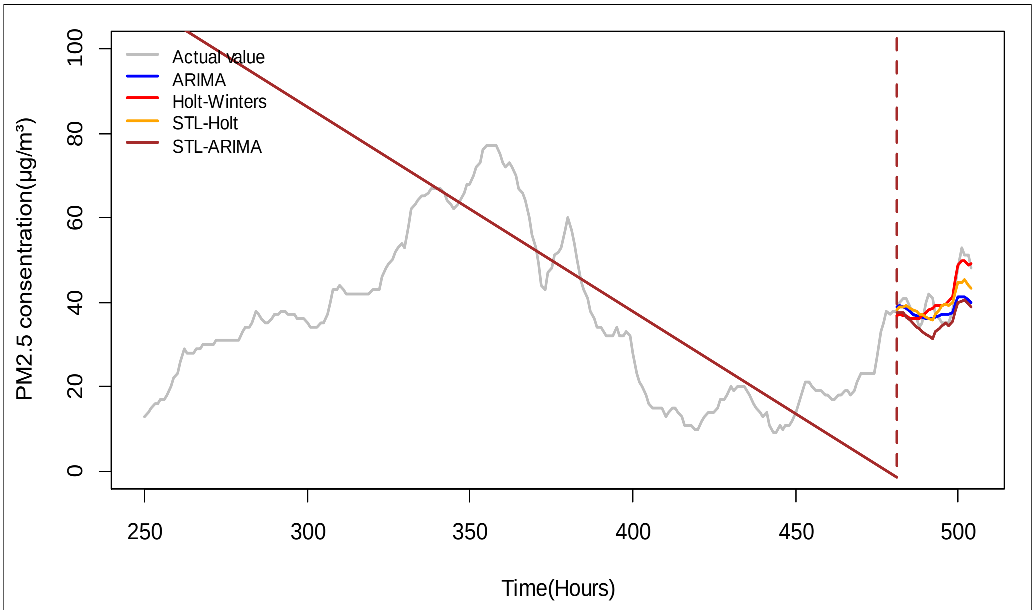

**Fig 3:** PM2.5 forecast result and measurement result line chart

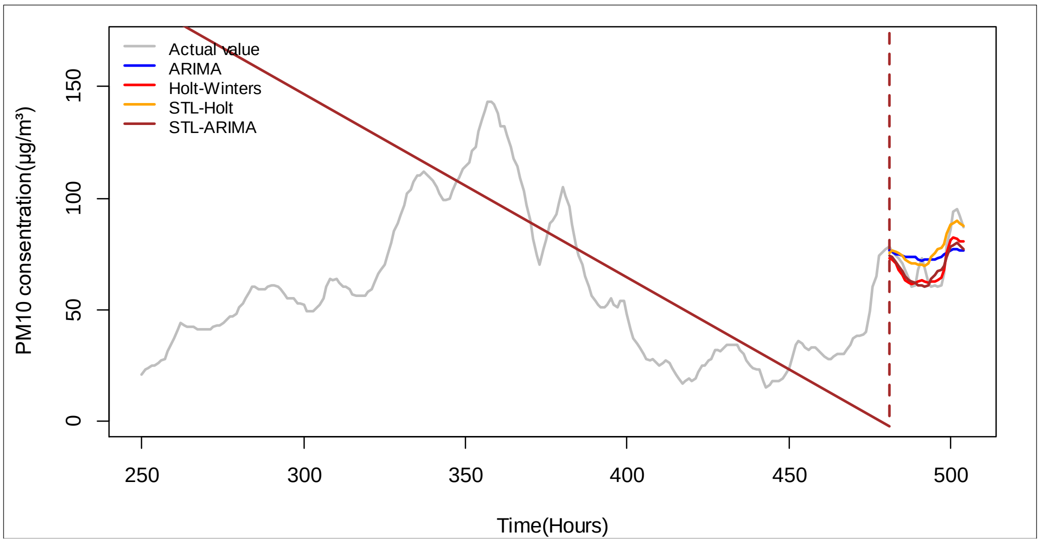

**Fig 4:** PM10 forecast result and measurement result line chart

### Comparison of model prediction accuracy

The performance results for the ARIMA model, Holt-Winters model and STL decomposition method are given in Table 5. In the prediction of PM2.5 concentrations, the Holt-Winters model had the smallest RMSE, MAE, and MAPE with values of 2.88, 2.55, and 6.47% respectively, followed by the STL- Holt method with 3.82, 3.19 and 7.63%. $R^2$ of the predicted values for the Holt-Winters model was the greatest with 0.73, and the SARIMA model was the next with 0.72. In the prediction of PM10 concentrations, the Holt-Winters model also had the smallest RMSE, MAE and MAPE, with 5.93, 4.94 and 6.44% in order, followed by the STL-ARIMA method with 7.40, 6.02 and 7.91%, respectively. The Holt-Winters model also had the highest $R^2$ of 0.89 for the predicted values, and the STL-ARIMA model was the next one with 0.75. Overall, the Holt-Winters model provides the best prediction of the two particulate matters, followed by the STL decomposition method.

**Table 5:** Prediction accuracy of models

| Models | PM2.5 | | | | PM10 | | | |
|---|---|---|---|---|---|---|---|---|
| | RMSE | $R^2$ | MAE | MAPE（%） | RMSE | $R^2$ | MAE | MAPE（%） |
| Sarima | 4.92 | 0.72 | 3.61 | 8.13 | 9.91 | 0.74 | 8.31 | 11.75 |
| Holt-Winters | 2.88 | 0.73 | 2.55 | 6.47 | 5.96 | 0.89 | 4.94 | 6.44 |
| STL-Holt | 3.82 | 0.66 | 3.19 | 7.63 | 8.27 | 0.69 | 6.41 | 9.79 |
| STL-ARIMA | 6.24 | 0.58 | 4.83 | 10.96 | 7.40 | 0.75 | 6.02 | 7.91 |

## 4. Conclusion

In this study, the ARIMA model, Holt-Winters model and STL decomposition were used to predict the hourly atmospheric particulate matter concentrations in Guangzhou for January 2022. The models were fitted with data from the first 20 days and predicted PM2.5 concentrations and PM10 concentrations for the next 24 hours.

The results showed that the four-time series methods have achieved favorable forecasting effects, and the predicted values obtained are consistent with the trend of the actual values. The four evaluation indexes RMSE, MAE, MAPE, $R^2$ of the Holt-Winters model were optimal, with values of 2.88, 2.55, 6.47%, 0.73 in the PM2.5 concentration prediction and 5.93, 4.94, 6.44%, 0.89 in the PM10 concentration prediction, respectively. Overall, the ARIMA model had the worst prediction performance, but the $R^2$ of it's prediction result was acceptable. The reason for this may be related to the application conditions of the model. The modelling process of the ARIMA model is more complex and requires a high degree of stationarity of the series data. The operation of difference transformation to make the series stationary may lose some information, thus reducing the predictive power of the model.

The time series forecasting method used in this study only uses historical data on atmospheric particulate matter concentrations to predict future trends, requiring only a single set of variable information, but achieving satisfactory forecasting results. In other studies, for example, Huang *et al.* (2021) [11] constructed random forest, XGBOOST, LSTM and gated recurrent unit network (GRU) models using other pollutant concentrations and meteorological data as predictor variables to predict the daily average PM2.5 and PM10 concentrations in Guangzhou, and the MAPE of the best prediction results obtained were 27.33% and 25.20%, respectively, while the MAPE values of the model in this paper are within 10%. Therefore, the forecasting method used in this paper is simpler and more practical than those used in other studies.

## References

1. Kim Y, Manley J, Radoias V. Medium-and long-term consequences of pollution on labor supply: evidence from Indonesia. IZA Journal of Labor Economics. 2017;6(1):1-15.
2. Miller KA, Siscovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, *et al*. Long-term exposure to air pollution and incidence of cardiovascular events in women. New England Journal of Medicine. 2007;356(5):447-458.
3. Simiyu LM, Chelule JC, Ayubu AO, Imboga H. Time series analysis of air pollution trends in Kenya using environmental Kuznets curve. International Journal of Statistics and Applied Mathematics. 2021;6(1):128-133.
4. Bhatti UA, Yan Y, Zhou M, Ali S, Hussain A, Qingsong H, *et al*. Time series analysis and forecasting of air pollution particulate matter (PM 2.5): An SARIMA and factor analysis approach. IEEE Access. 2021;9:41019-41031.
5. Ventura LMB, de Oliveira Pinto F, Soares LM, Luna AS, Gioda A. Forecast of daily PM 2.5 concentrations applying artificial neural networks and Holt–Winters models. Air Quality, Atmosphere & Health. 2019;12:317-325.
6. Wongrin W, Chaisee K, Suphawan K. Comparison of Statistical and Deep Learning Methods for Forecasting PM 2.5 Concentration in Northern Thailand. Polish Journal of Environmental Studies. 2023;32(2):1419-1431.
7. Aladağ E. Forecasting of particulate matter with a hybrid ARIMA model based on wavelet transformation and seasonal adjustment. Urban Climate. 2021;39:100930.
8. Cryer JD, Chan K. Time Series Analysis with Applications in R. 2nd Ed. New York: Springer-Verlag; c2008. p. 90-92.
9. Hyndman RJ, Athanasopoulos G. Forecasting: Principles and practice. Melbourne, Australia: O Texts; c2018.
10. Cleveland RB, Cleveland WS, Mcrae JE, Terpenning I. STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics. 1990;6(1):3-73.
11. Huang C, Fan DP, Lu J, Liao Q. Prediction of pm 2.5 and pm 10 concentration in Guangzhou based on deep learning model. Environmental Engineering. 2021;39(12):135-40.