Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Forecasting of Beijing $PM_{2.5}$ with a hybrid ARIMA model based on integrated AIC and improved GS fixed-order methods and seasonal decomposition

Lingxiao Zhao[a], Zhiyang Li[b], Leilei Qu[c,*]

[a] College of Ocean and Civil Engineering, Dalian Ocean University, Dalian 116024, China
[b] College of Civil Engineering, Chongqing University, Chongqing 400044, China
[c] College of Information Engineering, Dalian Ocean University, Dalian 116024, China

ARTICLE INFO

Keywords:
Integrated fixed-order
Autoregressive integrated moving average (ARIMA)
Beijing $PM_{2.5}$ prediction
Comprehensive evaluation coefficient (CEC)
Seasonal decomposition and reconstruction

ABSTRACT

Accurate particulate matter 2.5 ($PM_{2.5}$) prediction plays a crucial role in the accurate management of air pollution and prevention of respiratory diseases. However, $PM_{2.5}$, as a nonlinear time series with great volatility, is difficult to achieve accurate prediction. In this paper, a hybrid autoregressive integrated moving average (ARIMA) model is proposed based on the Augmented Dickey-Fuller test (ADF root test) of annual $PM_{2.5}$ data, thus demonstrating the necessity of first-order difference. The new method of using integrated akaike information criterion (AIC) and improved grid search (GS) methods is proposed to avoid the bias caused by using AIC alone to determine the order because the data are not exactly normally distributed. The comprehensive evaluation coefficient (CEC) is used to select the optimal parameter structure of the prediction model by considering multiple evaluation perspectives. The entropy value of the decomposed series is obtained by using range entropy A (RangeEn_A), and the series is reconstructed according to the entropy value, and finally the reconstructed series is predicted. We used Beijing $PM_{2.5}$ data for validation and the results showed that the new hybrid ARIMA model improved values of RMSE 99.23%, MAE 99.20%, R^2 118.61%, TIC 99.28%, NMAE 98.71%, NMSE 99.97%, OPC 43.13%, MOPC 98.43% and CEC 99.25% compared with the traditional ARIMA model. The results show that the method does greatly improve the prediction performance and provides a convincing tool for policy formulation and governance.

1. Introduction

Air pollution is one of the major problems threatening public health at this stage. A report by the World Health Organization (WHO) states that (World Health, 2016; World Health Organization. Regional Office for, 2002, 2003): More and more national air quality monitoring networks are measuring and monitoring $PM_{2.5}$, which reflects that the number of urban air pollution events are increasing worldwide, and there is a growing awareness of the health effects at the same time. Approximately more than 80% of citizens in cities where the air quality environment is monitored are susceptible to air quality levels that exceed the WHO guideline limit values.

In recent years, urban air quality levels have received widespread attention, and there are increasing calls to treat and control air pollutants such as particulate matter (PM), sulfur dioxide (SO_2), nitrogen oxides (NO_X), ozone (O_3), and carbon monoxide (CO) (Dai et al., 2019; Tang et al., 2019). It is no exaggeration to say that the development of heavy industries, the massive consumption of fossil fuels, the increase in urbanization, the dramatic increase in private transport ownership, agricultural activities such as straw burning, and several other activities have contributed directly or indirectly to the global air pollution problem (Guo et al., 2021; Vohra et al., 2021; Wen et al., 2020). The Chinese government also attaches great importance to the prevention and control of air pollution. Great efforts have been made in controlling pollutant emissions and dust control, promoting the construction of desulfurization and denitrification, developing green transportation, and strengthening motor vehicle exhaust emission control (Huang et al., 2019; Zhang et al., 2020).

Particulate matter (PM) has received more attention as the most hazardous component of air pollution. Among them, respirable particulate matter (PM_{10}) refers to particulate matter suspended in the air with an aerodynamic equivalent diameter ≤10 μm (Chen et al., 2017). In

* Corresponding author.
E-mail address: quleilei@dlou.edu.cn (L. Qu).

https://doi.org/10.1016/j.heliyon.2022.e12239
Received 24 July 2022; Received in revised form 17 November 2022; Accepted 1 December 2022

contrast, fine particulate matter ($PM_{2.5}$) refers to particulate matter with an aerodynamic equivalent diameter less than or equal to 2.5 μm in ambient air, which can be suspended in the air for a longer period of time (Cheng et al., 2019). Compared with PM_{10}, $PM_{2.5}$ has a smaller particle size, larger area, stronger activity, and is easily accompanied by toxic and harmful substances, and has a longer residence time and long transport distance in the atmosphere, thus having a greater impact on human health and atmospheric environmental quality (Cobbold et al., 2022; Zhou et al., 2021). Many scholars and experts have also contributed to air pollution control and haze management by finding that: atmospheric stability, geographical structure and meteorological conditions have significant effects on the distribution of aerosols in the air (Feng et al., 2015; Yang et al., 2022a; Zhang et al., 2020). $PM_{2.5}$ concentration prediction is currently recognized at home and abroad as the most critical pollutant for determining $PM_{2.5}$ concentration prediction is currently recognized as the most critical pollutant for determining urban air quality in China and abroad.

Meanwhile, during the 2019 coronavirus disease pandemic, Sharma et al. performed a statistical analysis through particulate matter concentrations measured in urban centers recorded during the quarantine period. The results revealed a significant reduction in harmful airborne particulate matter and a general improvement in air quality (Sharma et al., 2020). This study strongly illustrates the close relationship between air quality, especially $PM_{2.5}$, and human activities, but also the difficulty of prediction due to the uncertainty of human activities. Therefore, creating accurate and concise mathematical models to deal with different $PM_{2.5}$ series variations and fluctuation states in different study areas will help to accurately predict $PM_{2.5}$ quality.

Many different mathematical and statistical and mathematical models predict air pollution particulate matter. These prediction methods for linear and nonlinear data include: weather research and forecasting (WRF) model (Kong et al., 2021; Yang et al., 2021); regional atmospheric environment modeling system (RegAEMS) (Wang et al., 2012); extreme learning machine (ELM) model (Du et al., 2020; Wang et al., 2017a); artificial neural networks (ANN) (Biancofiore et al., 2017; Catalano et al., 2016); long short-term memory (LSTM) (Wen et al., 2019; Zhao et al., 2019); support vector machine (SVM) (Suárez Sánchez et al., 2011; Tian et al., 2022); recurrent neural network (RNN) (Belavadi et al., 2020); autoregressive integrated moving average model (ARIMA) (Theerthagiri, 2022; Zafra et al., 2017; Zhang et al., 2018).

Since the regression integrated moving average model was first proposed by Box and Jenkins in 1976, it has become a very mature theory of time series forecasting models after decades of development. It uses the existing prior values in the time series to try to forecast the future of the series (Box et al., 2015). Typically, time series on daily, monthly and quarterly scales are generated from components involving certain trends and seasonal variations. Numerous researchers have carried out predictive forecasting and analytical work on various types of complex time series using ARIMA models. Ning studied sediments with non-homogeneity, complex flow paths and fluid phase behavior to predict the production of unconventional reservoirs. The study data starts with the representative oil production data from a well located in an unconventional reservoir in the Denver-Julesburg (DJ) Basin. The results show that ARIMA and LSTM outperform Prophet. Also ARIMA is robust in predicting the oil production rate of wells across the DJ Basin (Ning et al., 2022). Sun constructed multiple historical ARIMA models using publicly available 2019 coronavirus disease data from Alberta, Canada. A method to modify the ARIMA model to accommodate heteroscedasticity time series was proposed by calculating the mean of the differences between predicted and corresponding actual values and their 95% confidence intervals (Sun, 2021). Kärner used ARIMA to compare the long-term temporal variability of top-of-atmosphere total solar irradiance (TSI) and surface air temperature series, demonstrating the dependence of various climate series on short-term fluctuations in TSI (Kärner, 2009). Arora and Keshari used a combination of the adaptive neuro-fuzzy inference system (ANFIS) and the ARIMA model to obtain reaeration coefficients that measure the interaction between the air-water interface at each sampling location in the river. The results showed significant improvement of the integrated ANFIS-ARIMA model in predicting the reaeration coefficients (Arora and Keshari, 2021). These studies demonstrate the effectiveness of ARIMA models in improving the accuracy of time series forecasting.

In order to continuously improve the prediction accuracy, some scholars have tried to apply multiple algorithms to form a hybrid ARIMA model for $PM_{2.5}$ time series prediction. Aladağ performed monthly prediction of PM_{10} concentration in Erzurum, Turkey, using appropriate coefficients selected by ARIMA for modeling. The results proved that the hybrid WT-ARIMA model based on wavelet transform has accurate prediction ability than the traditional ARIMA model for particulate pollution (Aladağ, 2021). Wang proposed a new hybrid Garch method integrating ARIMA and SVM for individual forecasting models with more reliable and accurate forecasting capability (Wang et al., 2017b).

However, most of the current research has focused on how to form a hybrid prediction model by multiple algorithms, and there is no discussion on the effectiveness and applicability of the model sizing method of the ARIMA model itself, and very few scholars have provided a suitable sizing method for the ARIMA model that takes into account the simplicity and accuracy of the model.

This study is an analysis of daily 1-km $PM_{2.5}$ data in China for 2018 estimated using an adaptive time modeling framework of satellite data and ground monitoring measurements (He et al., 2021), and the results are shown in Figure 1. The analysis found that $PM_{2.5}$ pollution is more serious in Xinjiang as well as in the Beijing-Tianjin-Hebei region, which is the capital economic circle of China and has an important political and economic status. And Beijing is located in the core of the region, and the study is of great significance. Several scholars have conducted studies from the perspective of environmental assessment. For example, Wang et al. investigated the superimposed compound growth relationship between high moisture content of water vapor transport and $PM_{2.5}$ and O3 pollution (Wang et al., 2022). Dong elucidated the sustainable development of air quality in Beijing, and analyzed $PM_{2.5}$ exposure in terms of the multi-scale spatial and temporal characteristics of $PM_{2.5}$ concentration and exposure risk intensity (Dong et al., 2022). Yang et al. combined linear mixed effect (LME) and geographically weighted regression (GWR) models to propose a two-stage statistical regression model with 1 km spatial resolution of aerosol optical thickness (AOD), meteorological variables and land use parameters as predictors to predict daily near-surface $PM_{2.5}$ concentrations in the Beijing-Tianjin-Hebei region from 2013-2017 (Yang et al., 2022b). Therefore, the daily $PM_{2.5}$ data (2953) obtained by the Environmental Protection Administration for Beijing from January 1, 2014 to January 31, 2022 were chosen as the study object to create the model when 66% were used as the training set and 34% as the test set.

To the best of our knowledge, traditional time series models usually use Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) criterion to fix the order (Akhter et al., 2020), and there is no study that combines AIC and other algorithms to explore the ARIMA model's suitable order. As a complement to the previous studies, this paper proposes a new hybrid method, which includes a combination of AIC and grid search algorithm results for ARIMA model order fixing and a hybrid prediction method based on seasonal decomposition, for predicting $PM_{2.5}$ concentration changes in Beijing. In addition, a comprehensive evaluation coefficient (CEC) is proposed for a more reasonable and comprehensive assessment of the prediction performance of the developed model by combining various error analysis indicators. Finally, it is believed that this study can be used as a guide for implementing various environmental regulations to control particulate matter pollution and provide technical reference for production and governmental decision making, which is crucial for accurate air pollution control and prevention of respiratory diseases.

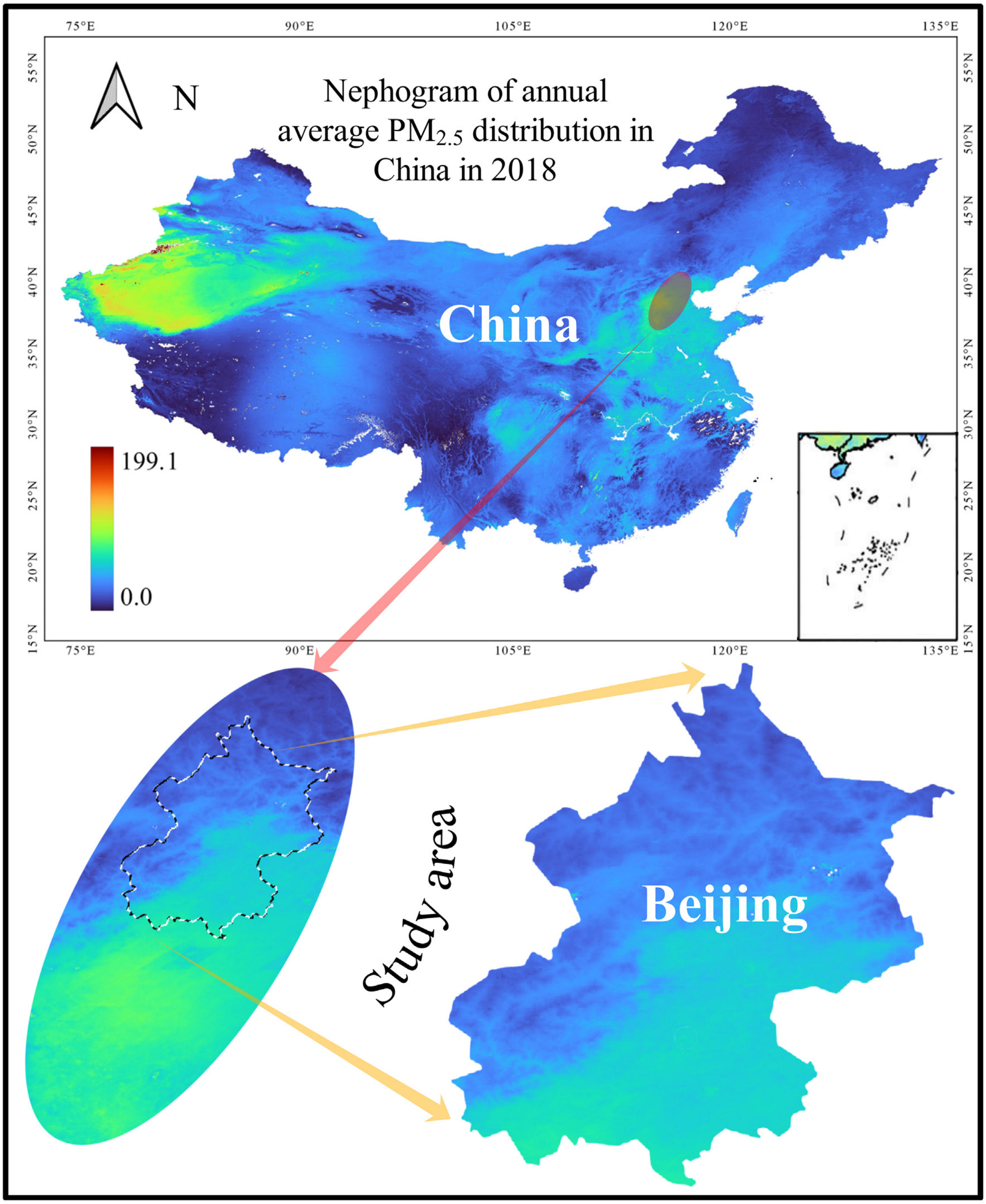


Figure 1. Location of Beijing in the study area and $PM_{2.5}$ distribution nephogram.

2. Methodology

2.1. Autoregressive integrated moving average model

2.1.1. Wold's decomposition and autocorrelation

The basic theorem of time series analysis, the Wold's decomposition, states that for each weakly smooth, purely nondeterministic stochastic process, $x_t - \mu$ can be written as a linear combination (or linear filter) of a series of uncorrelated random variables (Mills, 2019). “Purely nondeterministic” means that any deterministic component can be subtracted from $x_t - \mu$. For example some components can be perfectly predicted from their own past values. The linear filter is represented as follows [Eq. (1)].

$$x_t - \mu = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots = \sum_{j=0}^{\infty} \psi_j a_{t-j}, \psi_0 = 1 \tag{1}$$

where $a_t(t = 0, \pm1, \pm2, \ldots)$ are a series of uncorrelated random variables, which are drawn by Eq. (2) and Eq. (3) from the fixed distribution with:

$$E(a_t) = 0, V(a_t) = E(a_t^2) = \sigma^2 < \infty \tag{2}$$

$$Cov(a_t, a_{t-k}) = E(a_t, a_{t-k}) = 0, for\ all\ k \neq 0 \tag{3}$$

Such a sequence is called a white noise sequence, and can sometimes be innovatively represented as $a_t \sim WN(0,\sigma^2)$. The coefficients in Eq. (1) are called $\psi -$ *weights*.

Process of the model autocorrelation in x_t is described with Eq. (4) and Eq. (5) as follows :

$$E(x_t) = \mu \tag{4}$$

$$\gamma_0 = V(x_t) = E(x_t - \mu)^2 = \sigma^2 + \psi_1^2\sigma^2 + \psi_2^2\sigma^2 + \ldots = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 \tag{5}$$

By substituting the results of the white noise sequence $E(a_{t-i}a_{t-j}) = 0(i \neq j)$, we can obtain Eq. (6):

$$\gamma_k = E(x_t - \mu)(x_{t-k} - \mu) = \sigma^2(1 \cdot \psi_k + \psi_1\psi_{k+1} + \psi_2\psi_{k+2} + \ldots) = \sigma^2 \sum_{j=0}^{\infty} \psi_j\psi_{j+k} \tag{6}$$

Combining Eq. (5) and Eq. (6), which indicates Eq. (7).

$$\rho_k = \frac{\sum_{j=0}^{\infty} \psi_j\psi_{j+k}}{\sum_{j=0}^{\infty} \psi_j^2} \tag{7}$$

If $\psi -$ *weights* are assumed to be absolutely summable, then for an infinite number of weights there is $\sum_{j=0}^{\infty}|\psi_j| < \infty$, in which case the representation of the linear filter converges. The above condition is also equivalent to the assumption that x_t is smooth and allmoments are guaranteed to exist and are independent of time, especially the variance of x_t and γ_0 is finite.

2.1.2. Autoregressive model

Taking $\mu = 0$ without losing the generality of the problem, while letting $\psi_j = \phi^j$, then Eq. (1) can be written as [Eq. (8)]:

$$x_t = a_t + \phi(a_{t-1} + \phi a_{t-2} + \ldots) = \phi x_{t-1} + a_t \tag{8}$$

For Eq. (8) introducing the lag operator B, the lag expression can be described as [Eq. (9)]:

$$x_t = (1 - \phi B)^{-1} a_t = (1 + \phi \mathrm{B} + \phi^2 B^2 + \ldots) a_t \tag{9}$$

The above linear filter converges at $|\phi| < 1$ and is known as the stationarity condition.

2.1.3. Autoregressive integrated moving average model (ARIMA)

The inclusion of the nonsmooth autoregressive operator $\phi(B)$, $\phi(B) = 0$ with d unit roots like Eq. (10), in the previous model is effective for describing nonsmooth seasonal series.

$$\phi(B)\overline{x_t} = \theta(B)a_t = \phi(\mathrm{B})(1 - \mathrm{B})^d x_t \tag{10}$$

where $\phi(\mathrm{B})$ is the smooth autoregressive operator. Taking the difference operator $\nabla = 1 - B$ then for $d \geq 1, \nabla^d\overline{x_t} = \nabla^d x_t$ we have Eq. (11):

$$\theta(B)a_t = \phi(\mathrm{B})\nabla^d x_t = c(\mathrm{B})\omega_t \tag{11}$$

where $\omega_t = \nabla^d x_t$.

Reversing Eq. (11) yields can give Eq. (12).

$$x_t = S^d \omega_t \tag{12}$$

where S is defined according to Eq. (13) and Eq. (14) by the following infinite sum operator.

$$Sx_t = \sum_{h=-\infty}^{t} x_{\langle} = (1 + B + B^2 + \ldots)x_t = (1 - B)^{-1}x_t = \nabla^{-1}x_t \tag{13}$$

$$S^2x_t = Sx_t + Sx_{t-1} + Sx_{t-2} + \ldots = \sum_{I=-\infty}^{t}\sum_{h=-\infty}^{t} x_h = (1 + 2B + 3B^2 + \ldots)x_t \tag{14}$$

Since the infinite sum operator $S = (1 - B)^{-1}$ involved in the infinite sum does not converge, it cannot be used to define a nonsmooth ARIMA process in a practical process. A reasonable solution is to consider an infinite sum operator instead (Box et al., 2015), and for any positive integer m, the infinite sum operator S_m , $S_m^{(2)}$ can be expressed as Eq. (15) and Eq. (16):

$$S_m = (1 + B + B^2 + \ldots + B^{m-1}) = \frac{1 - B_m}{1 - B} \tag{15}$$

$$S_m^{(2)} = \sum_{j=0}^{m-1}\sum_{i=j}^{m-1} B^i = (1 + 2B + 3B^2 + \ldots + mB^{m-1}) = \frac{1 - B^m - mB^m(1 - B)}{(1 - B)^2} \tag{16}$$

Thus the relationship between the first-order differential ARIMA model and its corresponding smooth ARMA process can be expressed in terms of values up to a certain initial time point $k < t$ in the past, such as Eq. (17).

$$x_t = \frac{S_{t-k}}{1 - B^{t-k}}\omega_t = \frac{1}{1 - B^{t-k}}(\omega_t + \omega_{t-1} + \omega_{k+1}) \tag{17}$$

2.2. Seasonal decomposition

The basic theorem of time series analysis called Wold's decomposition, states that every weakly smooth, purely uncertain stochastic process can be written as a linear combination of a sequence of uncorrelated random variables (Mills, 2019). So seasonal decomposition models are divided into additive model and multiplicative model , the additive and multiplicative model can be described separately as follows [Eq. (18) and Eq. (19)] :

$$Y[t] = T[t] + S[t] + R[t] \tag{18}$$

$$Y[t] = T[t]^*S[t]^*R[t] \tag{19}$$

where $T[t]$ is the trend term, $S[t]$ is the seasonal term, and $R[t]$ is the residual term. The idea of additive decomposition model and multiplicative decomposition model is similar, and the following is an example of additive model, which is divided into 3 parts.

2.2.1. Trend item decomposition

The trend term is decomposed using the centralized moving mean method, and expressed by Eq. (20) separately for both cases of odd and even time series frequencies f.

$$T_t = \begin{cases} \dfrac{x_{t-\left(\frac{f-1}{2}\right)} + x_{t-\left(\frac{f-1}{2}\right)+1} + \ldots + x_{t+\left(\frac{f-1}{2}\right)-1} + x_{t+\left(\frac{f-1}{2}\right)}}{f}, f = 2k+1, t \in \left(\dfrac{f+1}{2}, l - \dfrac{f-1}{2}\right), k \in Z \\ \dfrac{0.5x_{t-\left(\frac{f}{2}\right)} + x_{t-\left(\frac{f}{2}\right)+1} + \ldots + x_{t+\left(\frac{f}{2}\right)-1} + x_{t+\left(\frac{f}{2}\right)}}{f}, f = 2k, t \in \left(\dfrac{f}{2} + 1, l - \dfrac{f}{2}\right), k \in Z \end{cases} \tag{20}$$

where t is the trend term, f is the time frequency series rate, and l is the time series length.

2.2.2. Seasonal cycle term decomposition and residual

The original time series is used to subtract the trend term, and the values at the same frequency in each period are averaged to obtain the seasonal term St_i, which can be represented by Eq. (21) and Eq. (22).

$$S_i = x_i - T_i \tag{21}$$

$$St_i = \frac{\sum_{i=0}^{n} S_{1+i*f}}{f}, t\in(1,f), n=\max\,(n, nf\le l) \tag{22}$$

Center the seasonal term, and the centered seasonal term St is described by Eq. (23).

$$S[t] = St - \overline{St} \tag{23}$$

The resulting residual term is given as follows [Eq. (24)].

$$R[t] = Y[t] - T[t] + S[t] \tag{24}$$

2.3. Grid search (GS) algorithm

The GS algorithm is an algorithm that exhausts the specified parameters and obtains the optimal parameters by cross-validating the parameters in the evaluation function (Abbaszadeh et al., 2022). The algorithm arranges the parameters into combinations to form a grid and calculates the corresponding parameters by traversal to obtain the optimal combination of parameters (Chang et al., 2022).

3. The hybrid ARIMA model

3.1. Sequence stationarity testing

To check the stability, an important feature of a time series, the augmented Dickey-Fuller (ADF) unit root test created with laged values of a series is used (Kębłowski and Welfe, 2004). An autoregressive process is called a unit root when the coefficient of the lag term is 1. When a

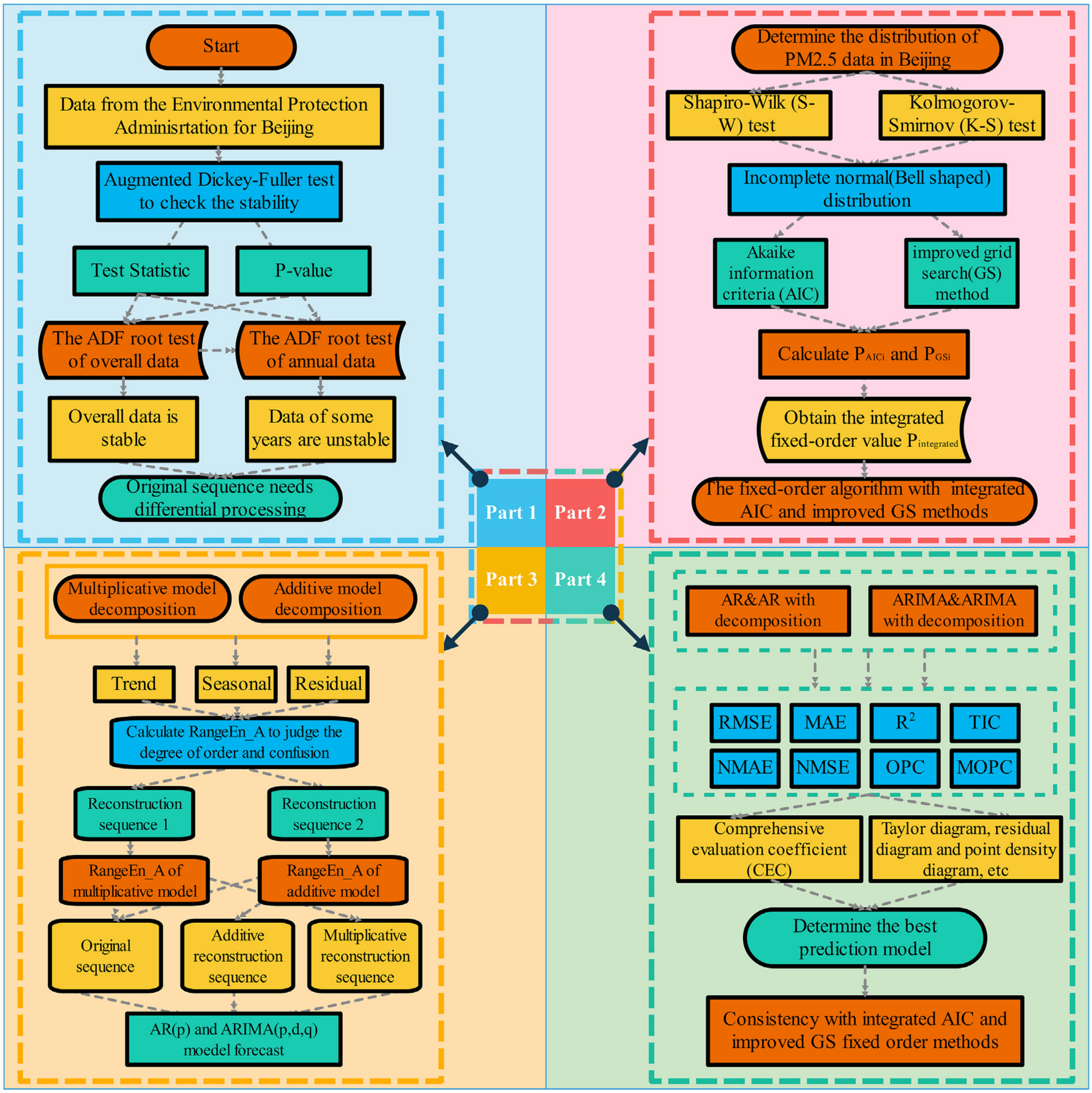

Start
Data from the Environmental Protection Adminisrtation for Beijing
Augmented Dickey-Fuller test to check the stability
Test Statistic
P-value
The ADF root test of overall data
The ADF root test of annual data
Overall data is stable
Data of some years are unstable
Original sequence needs differential processing
Determine the distribution of PM2.5 data in Beijing
Shapiro-Wilk (S-W) test
Kolmogorov-Smirnov (K-S) test
Incomplete normal(Bell shaped) distribution
Akaike information criteria (AIC)
improved grid search(GS) method
Calculate P_{AICi} and P_{GSi}
Obtain the integrated fixed-order value $P_{integrated}$
The fixed-order algorithm with integrated AIC and improved GS methods
Part 1
Part 2
Part 3
Part 4
Multiplicative model decomposition
Additive model decomposition
Trend
Seasonal
Residual
Calculate RangeEn_A to judge the degree of order and confusion
Reconstruction sequence 1
Reconstruction sequence 2
RangeEn_A of multiplicative model
RangeEn_A of additive model
Original sequence
Additive reconstruction sequence
Multiplicative reconstruction sequence
AR(p) and ARIMA(p,d,q) moedel forecast
AR&AR with decomposition
ARIMA&ARIMA with decomposition
RMSE
MAE
R^2
TIC
NMAE
NMSE
OPC
MOPC
Comprehensive evaluation coefficient (CEC)
Taylor diagram, residual diagram and point density diagram, etc
Determine the best prediction model
Consistency with integrated AIC and improved GS fixed order methods

Figure 2. The flowchart of the hybrid ARIMA prediction algorithm.

unit root exists, any error in the residual series does not decay with increasing sample size and the relationship between the independent and dependent variables is deceptive. Such a regression is also called a pseudo-regression, where the effect of residuals in the model is permanent and the process is random walk (Matsumoto et al., 2011). And the ADF root test can be used to determine the existence of unit root. The original hypothesis H0 has a unit root, and if the significance test statistic obtained is less than three confidence values of (10%, 5%, 1%), then the original hypothesis is rejected with (90%,95%,99%) certainty relative to (Bisaglia and Procidano, 2002). If the series is smooth then there is no unit root, otherwise there is a unit root.

In ARIMA (p,d,q), the traditional approach for model sizing is to consider the autocorrelation function (ACF) and partial autocorrelation function (PACF) to select the p and q values (Zhou et al., 2022). And then, based on the determined parameters, the Akaike information criteria (AIC) is used to determine the order from both prediction accuracy and model simplicity (Snipes and Taylor, 2014; Taddy, 2019). Where a smaller AIC value indicates better overall predictive performance of the model. The formula for calculating the AIC of the model is described as follows.

$$AIC = 2k - 2\ln\left(\widehat{L}\right) \tag{25}$$

where k is the number of estimated parameters in the model, $\widehat{L}$ is the maximum value of the likelihood function for the corresponding ARIMA model. The first part of Figure 2 shows the flow chart of sequence stationarity testing.

3.2. Integrated AIC and improved GS fixed-order methods

Normal distribution has a wide range of applications in production life and scientific experiments, and the probability distribution of many random variables can be described by normal distribution. The current Akaike information criteria (AIC) fixing process for finding the maximum likelihood estimate is mostly done by defaulting the probability density function of the data to the probability density function of the normal distribution, thus listing the log-likelihood function and finally finding the AIC value (Çankaya and Korbel, 2018). In contrast, the traditional AIC fixed-order method is implemented by $AIC = 2k - 2\ln(\widehat{L})$, where $\widehat{L}$ is maximum value of the likelihood function for the corresponding ARIMA model. Therefore, the use of AIC for ARIMA model sizing is not entirely accurate and appropriate when the data are not exactly normally distributed.

In order to determine the distribution of $PM_{2.5}$ data in Beijing, the Shapiro-Wilk (S–W) test (Zeng et al., 2019) and the Kolmogorov-Smirnov (K–S) test (Zhang and Cheng, 2004) were used. The results of the data normality test showed that the S–W test statistic was 0.782 and the K–S test statistic was 0.162, with a significance P-value of 0.000***, which was significant, so the original hypothesis was rejected, i.e., the data did not satisfy a perfectly normal distribution. However, the normal graph of Beijing $PM_{2.5}$ data shows a bell shape (high in the middle and low at both ends), which indicates that the data are not absolutely normal, but basically acceptable as a normal distribution. Thus, the premise of using the normal distribution probability density function (PDF) to find the maximum likelihood estimate and then using this value to find the AIC for model sizing is basically satisfied.

However, as mentioned above, since the data do not exactly conform to the normal distribution, the AIC order method based on the normal distribution of the data has some errors. Therefore, this paper proposes an integrated AIC and improved GS fixed-order methods. Improved grid search (GS) algorithm is based on the ARIMA traversal parameters (p,d,q) and adds the MSE of the posterior value when the traversal in order is greater than the previous value, the group traversal of this parameter is terminated and the next group traversal is performed. This improved GS algorithm can reduce the computational effort and time to some extent without affecting the accuracy of the results.

The specific process of the fixed-order algorithm with integrated AIC and improved GS methods is as follows: (1) calculate the fixed-order results using the AIC criterion and GS search, respectively; (2) subtract a constant term H_0 from each value in the calculated result series, where $H_0 = \prod(min_{i=1}^{n}(\mathrm{H}_i))$. This step can simplify the calculation and reduce the error, and also does not result in too small coefficient of variation (C_v) (Zhan et al., 2022); (3) Let $m = H_i - H_0(i = 1, 2, \ldots, n)n = H_{max} - H_{min}$ to obtain $P_i = \frac{m}{n} \times 100\%$. From this, the values of P_{AICi} and P_{GSi} for each order of ARIMA model are obtained; (4) the integrated fixed-order value $P_{integrated}$ is obtained, where $P_{integrated} = \alpha \times P_{AICi} + P_{GSi}$, where is the integrated regulator of Beijing $PM_{2.5}$ sequence. Part 2 of Figure 2 shows the flow chart of fixed-order methods with ARIMA model.

3.3. Sequence seasonal decomposition and reconstruction

The Beijing $PM_{2.5}$ series data were decomposed into trend, seasonal, and residual terms using the traditional seasonal decomposition method. And after that, the complexity and confusion degree of each sub-item were calculated using RangeEn_A (Omidvarnia et al., 2018). Based on the RangeEn_A values, the decomposed sequence subterms with similar RangeEn_A values are divided into two classes for additive and multiplicative reconfiguration to obtain two new sequences. Part 3 of Figure 2 shows the flow chart of the sequence seasonal decomposition and reconfiguration process.

3.4. Identifying the optimal individual model structure

Comprehensive evaluation coefficient (CEC) is based on different evaluation perspectives, including the accuracy of the predicted data compared with the original data, the predicted data, and whether the

Table 1. The error evaluation criteria for comparing ARIMA models with different parameter structures.

Criteria	Definition	Formula
RMSE	Root Mean Square Error	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i')^2}$
MAE	Mean Absolute Error	$\frac{1}{n}\sum_{i=1}^{n}\lvert Y_i - Y_i' \rvert$
R^2	Nash-Sutcliffe Efficiency Coefficient	$1 - \frac{\sum_{i=1}^{n}(Y_i - Y_i')^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2}$
TIC	Theil Inequality Coefficient	$\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i')^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i'^2}}$
NMAE	Normalized Mean Absolute Error	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Y_i - Y_i' \rvert}{\sqrt{Y_i \cdot Y_i'}}$
NMSE	Normalized Mean Square Error	$\frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - Y_i')^2}{Y_i \cdot Y_i'}$
OPC	Orientational Prediction Coefficient	$\frac{1}{n}\sum_{i=1}^{n}P_t$
MOPC	Median Orientational Prediction Coefficient	$\frac{1}{n}\sum_{i=1}^{n}(Q_i - Q_i')^2$
CEC	Comprehensive evaluation coefficient	0.1*(RMSE + MAE)+0.2*(TIC + NMAE + NMSE)+0.3*((1-OPC)+MDPC)+(1-R^2)

Notes: Y_i and Y_i' are the actual and predicted values of this time series in time period i. n is the number of testing datasets. p is the number of characteristic variables. $P_t = \begin{cases} 1, (Y_{i+1} - Y_i)(Y_{i+1}' - Y_i) \geq 0 \\ 0, \textit{otherwise} \end{cases}$, $Q_i = \begin{cases} 1, (Y_{i+1} - Y_i) \leq 0 \\ 0, \textit{otherwise} \end{cases}$, $Q_i' = \begin{cases} 1, Y_{i+1}' - Y_i') \leq 0 \\ 0, \textit{otherwise} \end{cases}$.

predicted data are consistent with the change in the same direction as the actual data. Multiple independent evaluation indexes are constructed in a weighted form to comprehensively evaluate the model's Beijing $PM_{2.5}$ prediction capability. Table 1 shows the model evaluation indicators and their calculation methods. their calculation methods.

As shown in Table 1, the accuracy between the actual and predicted values of a single model is determined by the comprehensive evaluation coefficient (CEC). The advantage of this index is that the CEC comprehensive evaluation coefficient is proposed by considering the error, fitting accuracy, and prediction direction through the single indexes of RMSE, MAE, R^2, TIC, NMAE, NMSE, OPC and MOPC. The last part of Figure 2 shows the process to identifying the optimal individual model structure.

4. Empirical study

In this section, we perform an example validation using the proposed method and the hybrid ARIMA model, computed using the Jupyter Notebook application, programmed in Python 3.9.0, and all experiments run on an NVIDIA GeForce GTX 1650 GPU.

4.1. Study area and available data

Beijing is located at the northwest edge of the North China Plain, between 39°28′ ~ 41°03′ N latitude and 115°25′ ~ 117°35′ E longitude. Since the 21st century, through the treatment of coal-fired boilers, the clean-up of civil fuels, industrial restructuring and other measures, Beijing has achieved remarkable results in the treatment of air pollution: the annual average concentration of sulfur dioxide dropped by 93.3%; in the past five years, the annual average concentration of $PM_{2.5}$ dropped from 89.5 $\mu g/cm^3$ in 2013 to 58 $\mu g/cm^3$ with a 35% decrease. In 2020, Beijing's annual average $PM_{2.5}$ concentration dropped to "30+" for the first time, at 38 $\mu g/cm^3$, down 57.5% compared to 2013. The number of heavily polluted days in Beijing decreases significantly, with 10 heavily polluted days in 2020, a decrease of 48 days or 82.8% from 2013.

4.2. Data description

In this paper, daily Beijing $PM_{2.5}$ data obtained from the Environmental Protection Administration. Beijing haze $PM_{2.5}$ (2953 in total) from January 1, 2014 to January 31, 2022 was selected as the dataset for the practical validation of the proposed hybrid ARIMA model. 66% of the collected data were classified as the training set and the remaining 34% were classified as the test set to validate the performance of the model. Figure 3 shows the characteristics of the Beijing $PM_{2.5}$ time series dataset.

4.3. Analysis of substantiation results

4.3.1. ADF unit root test for sequence stationarity

In this paper, we use Augmented Dickey-Fuller (ADF) root test to test the Beijing $PM_{2.5}$ time series year by year based on our unique perspective and find that although the ADF root test results for the overall data show smoothness, the annual year test results indicate that the data are not always smooth.

As shown in Table 2, the Test Statistic of the overall data is -5.183468, which is much smaller than the Critical Value (1%, 5%, 10%) of (−3.432608, -2.862538, -2.567301), while the p-value of 0.00001 is much smaller than 0.05. Therefore, the overall $PM_{2.5}$ sequence data is judged as a smooth sequence.

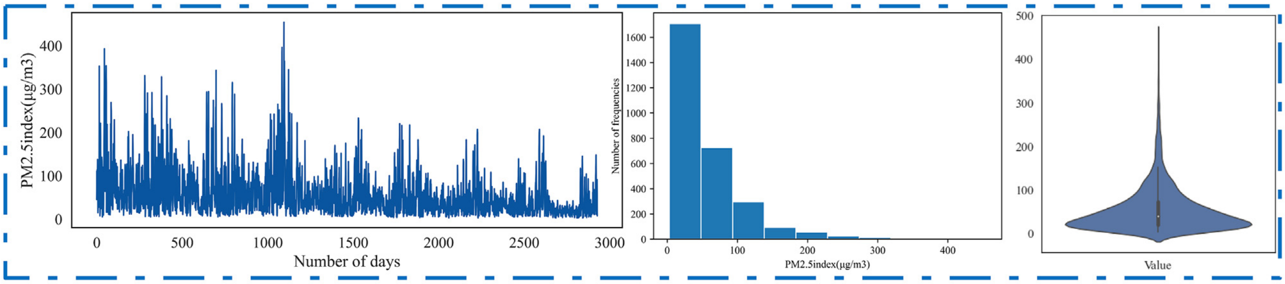
However, since the Beijing $PM_{2.5}$ time series has a clear temporal significance. The overall data has a seasonal frequency with a period of 365 days. Therefore, when conducting the ADF unit root test, the autolag is selected based on the AIC value, and the Lags used will be automatically selected in the ADF root test according to the characteristics of the $PM_{2.5}$ time series data in Beijing year by year, with a suitable lag term k. The results are shown in Table 3, the Test Statistic for 2017 is -1.58076, which is greater than the Critical Value (10%) of -2.57115, while the p-value is 0.493252 greater than 0.05, so the original hypothesis is rejected and the series is not smooth; the Test Statistic for 2021 is - 1.81452, which is greater than the Critical Value (10%) of -2.57147, while the p-value is 0.373280 is greater than 0.05, so the original hypothesis is rejected and the series is also unstable. Meanwhile, the Test Statistic for 2022 is -3.02662 which is smaller than the Critical Value (5%, 10%) (−2.96407, -2.62117), but larger than the Critical Value (1%) of -3.66992. Therefore, there is 99% probability to reject the original hypothesis and the data is still can be considered non-stationary.

Combining the above ADF root test results, first-order differencing of the data was considered to reduce the irregular fluctuations among the data.

4.3.2. Individual AIC and improved GS method results

The results of the Akaike information criteria (AIC) definite order are shown in Figure 4.

From the results in Figure 4, it can be seen that for the original undifferentiated series, the minimum AIC value is 29876.4 and the ARIMA final parameter sizing result is (8,0,1). For the first-order differential sequence, the minimum AIC value is 29857.5, and the ARIMA final parameter fixing order result is (8,1,2). Therefore, the AIC value of the series after the first-order differencing has a significant reduction of 18.9, which indicates that the differencing effectively reduces the volatility between the data and is beneficial to the accurate prediction of the ARIMA model. It also proves the necessity of the above year-by-year ADF root test for the data.



	count	mean	std	min	25%	50%	75%	max	skewness	kurtosis	mode
Beijing	2953	55.471	53.233	0	20	39	73	454	2.283511	7.350224	39

Figure 3. Details information of the Beijing area.

Table 2. ADF unit root test results of overall data.

Statistics	Value
Test Statistic	-5.183468
p-value	0.00001
#Lags Used	28
Number of Observations Used	2899
Critical Value (1%)	-3.432608
Critical Value (5%)	-2.862538
Critical Value (10%)	-2.567301

The smoothness test of the year-by-year $PM_{2.5}$ series in Beijing identified the unsteadiness of some of the data, and the first-order difference greatly reduced the large volatility of this part of the data, which is the reason why the AIC value of ARIMA (8,1,2) decreased by 18.9 compared with that of ARIMA (8,0,1).

Figure 5 shows the results of the improved GS method, and we plotted a four-dimensional image to show it. Where (x,y,z) represent ARIMA (p,d,q) respectively, and the color shades of the circles indicate the MSE values after searching by grid search. As for the untraversed parameter in the improved GS method, the MSE predicted by this parameter model is greater than the MSE that has been traversed, so it is not calculated to save time r arithmetic power.

From Figure 5 it can be seen that the final parameter fixing order result of the GS method ARIMA is (6,1,2) and the MSE value is 544.403.

4.3.3. Integrated AIC and improved GS order determination results

Following the algorithm proposed in Section 3.3, P_{AICi} and P_{GSi} for each order of ARIMA model are calculated as shown in Tables 4 and 5 below respectively.

From Table 4 $(P_{AICi})_{min} = 0.026\%$, the model order is (8,1,2); Table 5 $(P_{GSi})_{min} = 0.038\%$, the model order is (6,1,2); thus, we can find $P_{integrated}$, where $P_{integrated} = \alpha \times P_{AICi} + P_{GSi}$. The experimental test of Beijing $PM_{2.5}$ data from January 1, 2014 to January 31, 2022, $\alpha = 0.02$ is chosen as the integrated regulator of Beijing. Then the calculation results of $P_{integrated}$ are shown in Table 6.

From Table 6 $(P_{integrated})_{min} = 0.057\%$, so by this method, the final order of ARIMA model is fixed as (6,1,2).

4.3.4. Sequence seasonal decomposition and reconstruction results

The traditional time series seasonal decomposition is used, and the multiplicative and additive decomposition are performed separately for the original $PM_{2.5}$ in Beijing, while the frequency is specified as a quarter of time, i.e., 90 days is a frequency.

For the multiplicative and additive models decomposed trend, seasonal and residual calculated RangeEn_A (Omidvarnia et al., 2018). The results are shown in Table 7.

- As shown in Table 8. For the multiplicative and additive model decomposition, the RangeEn_A values of the trend terms are both 0.0263. Since for data prediction, the reconstructed sequence with the minimum RangeEn_A value is the key determinant of the $PM_{2.5}$ time series, the trend term is used as the reconstruction sequence 1

Table 3. ADF unit root test results of annual data.

	2014	2015	2016	2017	2018	2019	2020	2021	2022
Test Statistic	-6.52431	-10.64571	-11.36423	-1.58076	-10.74573	-10.80132	-10.01065	-1.81452	-3.02662
p-value	1.02E-08	4.80E-19	9.31E-21	0.493252	2.74E-19	2.01E-19	1.78E-17	0.373280	0.032482
#Lags Used	4	1	1	10	1	1	1	17	0
Number of Observations Used	360	363	363	353	358	363	364	329	30
Critical Value (1%)	-3.44865	-3.44849	-3.44849	-3.44901	-3.44875	-3.44849	-3.44844	-3.45038	-3.66992
Critical Value (5%)	-2.86960	-2.86954	-2.86954	-2.86976	-2.86965	-2.86954	-2.86951	-2.87037	-2.96407
Critical Value (10%)	-2.57107	-2.57103	-2.57103	-2.57115	-2.57109	-2.57103	-2.57102	-2.57147	-2.62117

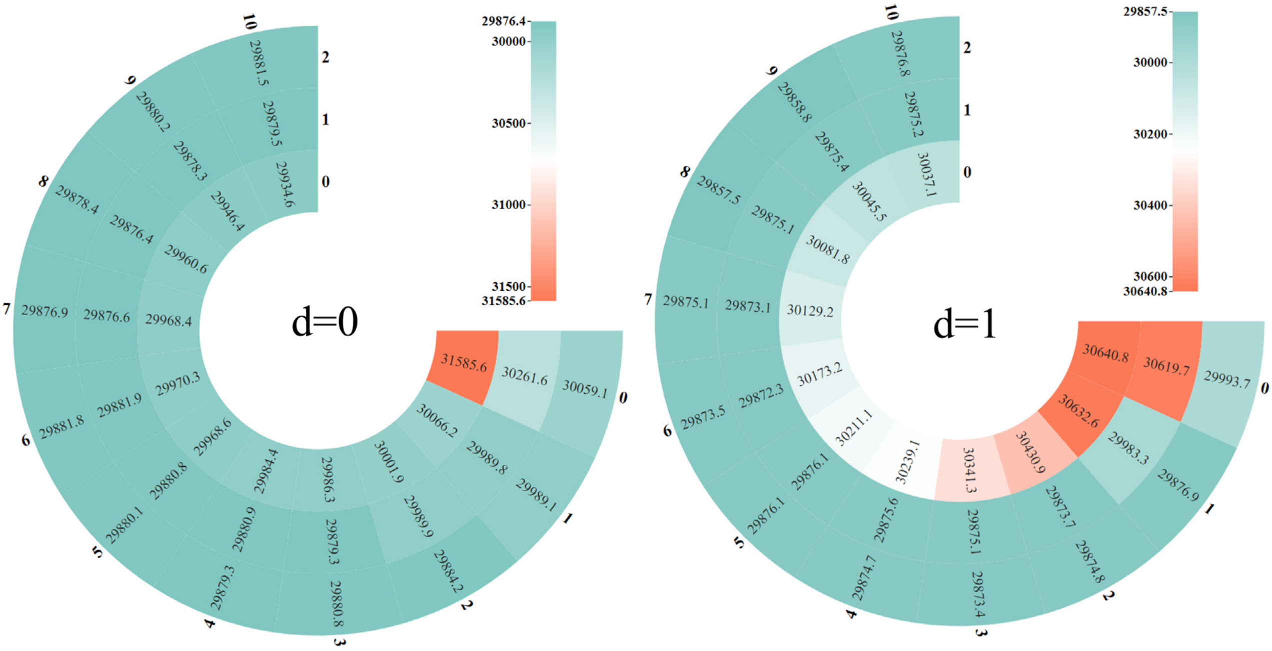


Figure 4. Results of AIC criterion with original sequence undifferentiated and first-order difference.

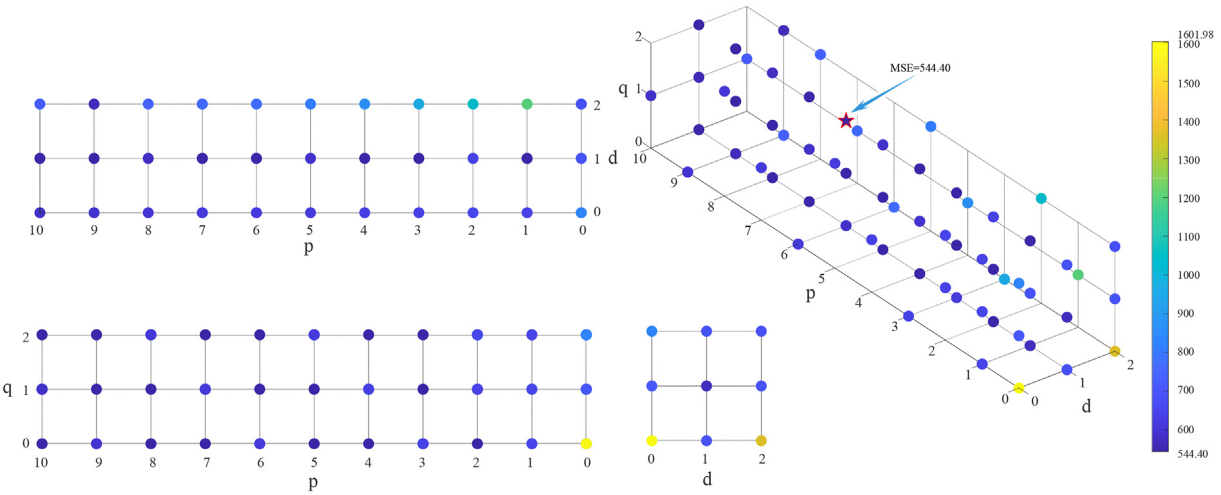


Figure 5. Results of MSE values for the GS method.

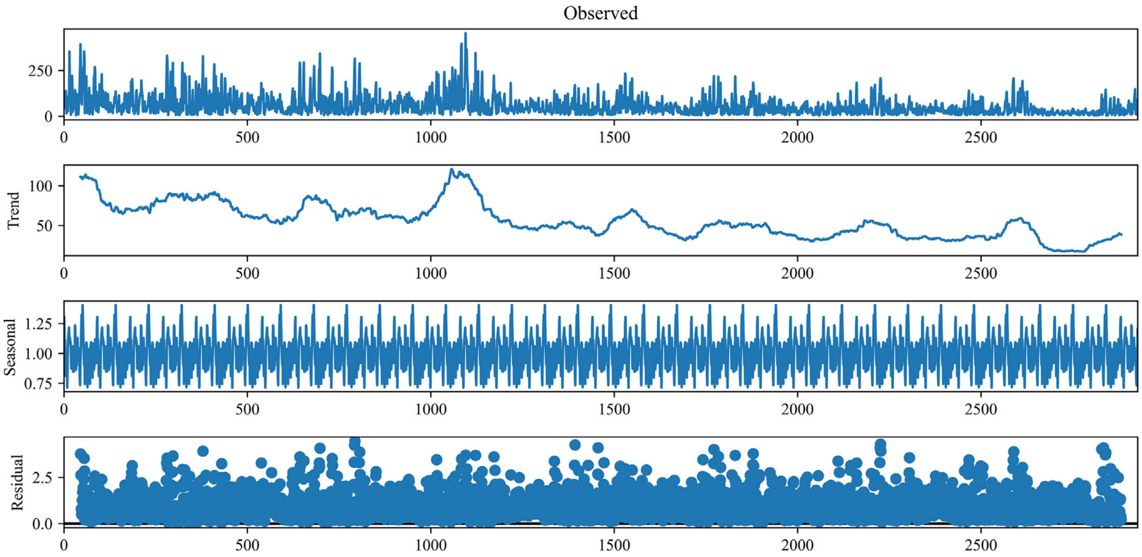


Figure 6. The seasonal decomposition of the multiplicative model results with 90 frequency.

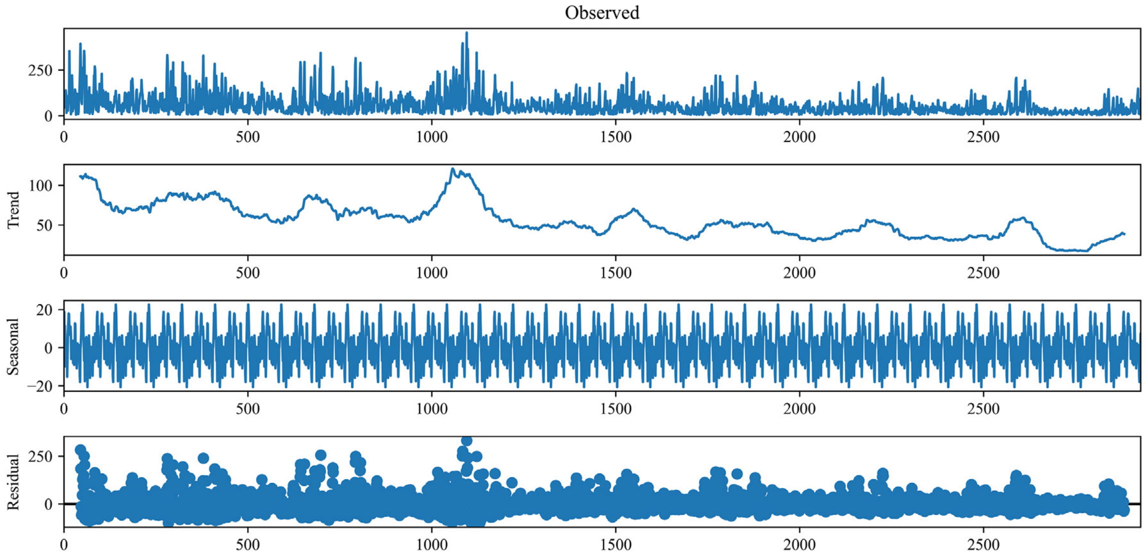


Figure 7. The seasonal decomposition of the additive model results with 90 frequency.

Table 4. P_{AICi} value for each order of ARIMA model.

AIC q\p	0	1	2	3	4	5	6	7	8	9	10
0	45.354%	44.878%	33.209%	28.023%	22.107%	20.493%	18.298%	15.753%	13.009%	10.905%	10.413%
1	44.136%	7.309%	0.966%	1.041%	1.074%	1.100%	0.887%	0.931%	1.046%	1.062%	1.055%
2	7.909%	1.149%	1.028%	0.946%	1.022%	1.099%	0.956%	1.047%	**0.026%**	0.105%	1.144%

Table 5. P_{GSi} value for each order of ARIMA model.

GS q\p	0	1	2	3	4	5	6	7	8	9	10
0	14.168%	14.539%	9.892%	8.134%	5.773%	5.216%	4.854%	4.842%	3.887%	2.977%	2.483%
1	15.054%	3.384%	0.343%	0.329%	0.300%	0.247%	0.171%	0.103%	0.166%	0.219%	0.220%
2	3.241%	0.390%	0.281%	0.187%	0.174%	0.160%	**0.038%**	0.077%	0.064%	0.147%	0.229%

Table 6. The results of $P_{integrated}$ value for integrated AIC and improved GS algorithm.

$P_{integrated}$ q\p	0	1	2	3	4	5	6	7	8	9	10
0	15.075%	15.437%	10.556%	8.695%	6.215%	5.626%	5.220%	5.157%	4.147%	3.195%	2.691%
1	15.936%	3.530%	0.363%	0.350%	0.321%	0.269%	0.189%	0.121%	0.187%	0.240%	0.241%
2	3.399%	0.413%	0.301%	0.206%	0.195%	0.182%	**0.057%**	0.098%	0.065%	0.149%	0.252%

and the seasonal and residual as the reconstruction sequence 2. The seasonal and noise terms are thus removed from the original data in reconstruction sequence 1, and only the low RangeEn_A values are retained. trend term.

- The Haze $PM_{2.5}$ sequence can not be separated from human production life. From the perspective of haze formation, polluting emissions from the operation of heavy industrial enterprises, exhaust emissions from automobiles, and dust from coal-fired power generation are all important causes of haze. The Chinese government attaches great importance to the management of the air environment. It is also easy to find from the observed and trend terms in Figures 6 and 7 that since the peak of $PM_{2.5}$ series index value in Beijing in 2016, it has been showing a fluctuating downward trend since then. It is very encouraging that the environment we live in is becoming better and better while taking into account the development.

4.3.5. Hybrid prediction model error test results

To test the proposed hybrid ARIMA model and integrated AIC and improved GS fixed-order methods, several performance evaluation metrics were compared, but judging the merit of the model cannot be determined simply based on a single evaluation metric. Here, the effectiveness of the model is chosen to be determined based on the minimum value of CEC, which includes RMSE, MAE, R^2, TIC, NMAE, NMSE, OPC and MOPC. Among them, RMSE and MAE are tested for extreme errors and prediction outliers; R^2 responds to the degree of fit of the prediction model; TIC, NMAE and NMSE determine the accuracy of the model built; OPC and MOPC ensure the consistency between the predicted and the actual direction of change of $PM_{2.5}$ series. The above indicators are combined by weighting to form the Comprehensive evaluation coefficient (CEC), which selects the best structure for each model.

The original ARIMA best model parameter structure and the corresponding performance evaluation coefficients are given in Table 9, with the AR model listed as a comparison. The performance evaluation metrics corresponding to the optimal model parameters for the hybrid ARIMA and AR models are given in Table 10.

Table 7. Range entropy (RangeEn_A) of all items via multiplicative and additive decomposition.

	Trend	Seasonal	Residual
Multiplicative model	0.0263	0.7974	0.8206
Additive model	0.0263	0.8314	0.7734

From Table 9, it can be seen that the highest prediction accuracy is achieved in the ARIMA model when the parameters are fixed-order ARIMA (6,1,2). In comparison, the highest prediction accuracy of the AR model is achieved with the parameter fixed order AR (5). However, the CEC value of ARIMA (6,1,2) is 5.0546 which is 0.0384 lower than that of AR (5) which is 5.0930, so ARIMA has higher prediction performance. This result also validates our proposed integrated AIC and improved GS fixed-order algorithm, and the result shows that it is consistent with the actual computational results.

For the AR and ARIMA models with the best prediction performance in Table 9, multiplicative and additive model decomposition and reconstitution. were performed for the reconstructed sequences, and the results are shown in Figure 6 and Figure 7. The hybrid ARIMA model has a substantially higher prediction performance compared to The prediction results of the Hybrid ARIMA model are substantially better than the original ARIMA model, but the results of the additive model are better than the multiplicative model.

Among all AR models, the AR (5)-additive model has the best prediction performance. Compared with the AR (5) model which has the best performance in the original model, the RMSE improves 97.91%, MAE improves 98.82%, R^2 improves 119.68%, TIC improves 98.06%, NMAE improves 98.30%, NMSE improves 99.86% CEC metrics improved by 98.45%.

Among all ARIMA models, the ARIMA (6,1,2)-add model has the best prediction performance. Compared with the ARIMA (6,1,2) model, which has the best performance in the original model, RMSE improves by 99.23%, MAE improves by 99.20%, R^2 improves by 118.61%, TIC improves by 99.28%, NMAE improves by ARIMA model has better performance than AR model in terms of prediction accuracy, directional prediction accuracy and model fit.

Table 8. Range entropy (RangeEn_A) of reconstructed sequence and its specific value.

	Serial number	Multiplicative model	Additive model
Reconstruction sequence	1	Trend	Trend
	2	Seasonal×Residual	Seasonal + Residual
RangeEn_A	1	0.0263	0.0263
	2	0.6543	1.6048

Table 9. Prediction results and calculation errors of original AR and ARIMA model.

Method	RMSE	MAE	R^2	TIC	NMAE	NMSE	DPC	MDPC	CEC
AR1	24.0701	16.7683	0.4226	0.2693	0.5576	0.5979	0.6546	0.4799	5.1938
AR2	25.1531	17.3064	0.3694	0.3100	0.6573	0.9634	0.5944	0.4528	5.5202
AR3	24.6325	16.9609	0.3953	0.2990	0.6144	0.7571	0.6185	0.4578	5.3500
AR4	24.9545	17.1675	0.3794	0.3057	0.6351	0.8794	0.6024	0.4608	5.4544
AR5	**23.3822**	**16.7117**	**0.4551**	**0.2629**	**0.5835**	**0.6677**	**0.6747**	**0.4608**	**5.0930**
AR6	24.3858	16.7992	0.4073	0.2941	0.6020	0.7083	0.6235	0.4669	5.2851
AR7	24.5058	16.8831	0.4015	0.2968	0.6148	0.7818	0.6175	0.4669	5.3309
AR8	24.3449	16.7657	0.4093	0.2931	0.5976	0.6926	0.6285	0.4649	5.2693
AR9	24.1695	16.6348	0.4178	0.2892	0.5860	0.6605	0.6365	0.4639	5.2180
ARIMA012	24.0473	17.1153	0.4237	0.2718	0.5821	0.6357	0.6898	0.4618	5.2221
ARIMA112	23.4120	16.6435	0.4537	0.2615	0.5644	0.6053	0.6918	0.4669	5.0706
ARIMA212	23.3873	16.6179	0.4549	0.2613	0.5662	0.6179	0.6817	0.4608	5.0685
ARIMA312	23.3662	16.6133	0.4558	0.2610	0.5654	0.6116	0.6898	0.4629	5.0616
ARIMA412	23.3633	16.6159	0.4560	0.2609	0.5658	0.6154	0.6837	0.4608	5.0635
ARIMA512	23.3601	16.6195	0.4561	0.2609	0.5661	0.6146	0.6847	0.4629	5.0636
ARIMA522	24.6440	17.2345	0.3947	0.2660	0.5983	0.7170	0.6245	0.4839	5.3672
ARIMA612	**23.3324**	**16.6021**	**0.4574**	**0.2605**	**0.5641**	**0.6057**	**0.6928**	**0.4679**	**5.0546**
ARIMA712	23.3413	16.6140	0.4570	0.2606	0.5647	0.6065	0.6888	0.4618	5.0568
ARIMA812	23.3383	16.6152	0.4571	0.2606	0.5648	0.6072	0.6898	0.4639	5.0570
ARIMA822	24.4062	17.1962	0.4063	0.2663	0.5827	0.6708	0.6406	0.4829	5.3106
ARIMA912	23.3571	16.6235	0.4563	0.2608	0.5648	0.6071	0.6918	0.4649	5.0603

Table 10. Prediction results and calculation errors of hybrid ARIMA model.

	RMSE	MAE	R^2	TIC	NMAE	NMSE	DPC	MDPC	CEC
AR5-mul	0.54774	0.18987	0.99970	0.00571	0.00537	0.00015	0.98738	0.00946	0.08293
ARIMA612-mul	0.21101	0.12583	0.99996	0.00220	0.00373	0.00002	0.99159	0.00736	0.03965
AR5-add	0.48789	0.19704	0.99976	0.00509	0.00990	0.00091	0.98738	0.01157	0.07916
ARIMA612-add	**0.17950**	**0.13244**	**0.99997**	**0.00187**	**0.00730**	**0.00019**	**0.99159**	**0.00736**	**0.03783**

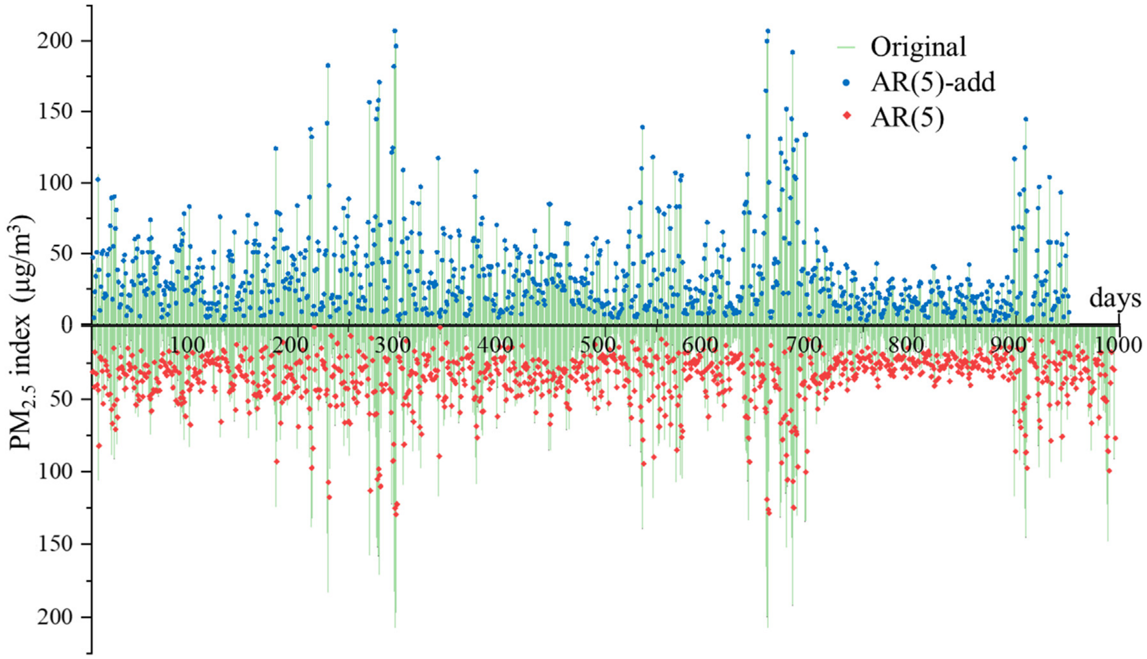


Figure 8. Comparison of AR (5)-add and AR (5) model predicted value and actual value.

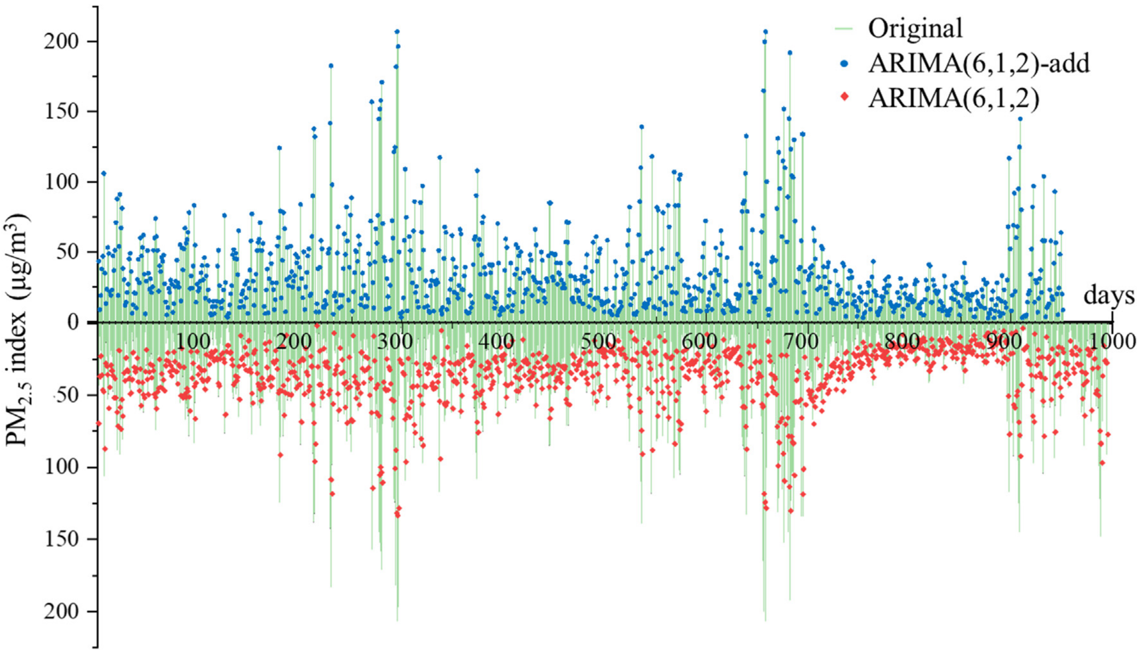


Figure 9. Comparison of ARIMA (6,1,2)-add and ARIMA (6,1,2) model predicted value and actual value.

4.4. Hybrid ARIMA predictive analysis

In the last section, the six models with better prediction performance, AR (5), AR (5)-mul and AR (5)-add model; ARIMA (6,1,2), ARIMA (6,1,2)-mul and ARIMA (6,1,2)-add, are focused on. For the above 6 methods plotted images to explore the prediction accuracy and the variation of the prediction residuals of the whole $PM_{2.5}$ time series. And the following 2 conclusions were obtained.

(1) As in subsection 4.3.5, we selected AR (5)-add, ARIMA (6,1,2)-add with the best prediction performance to compare with AR (5), ARIMA (6,1,2) with the best original prediction, where since the seasonal decomposition was set at a frequency of 90 for one quarter, the prediction accuracy after The results are shown in Figures 8 and 9.

The results show that for both the AR (5)-add and ARIMA (6,1,2)-add models, the test set fits better. However, for the AR (5) and ARIMA (6,1,2) models, the fit of the test set is slightly worse. Among them, the worst prediction is done for the days when $PM_{2.5}$ is 125 or higher concentration, which indicates that the traditional AR and ARIMA models cannot complete a good fit and prediction for the time series with large volatility and unusually extreme (very large or very small) series data. Both Figures 8 and 9 present such characteristics, but with the additive model, it is able to solve these two problems well and perform better regression and prediction of the series.

(2) The residuals of the six models with good prediction performance are plotted for the whole process of the test set, and the results are shown in Figure 10. It is obvious that the residuals of the traditional AR and ARIMA model are about 60 times more than those of

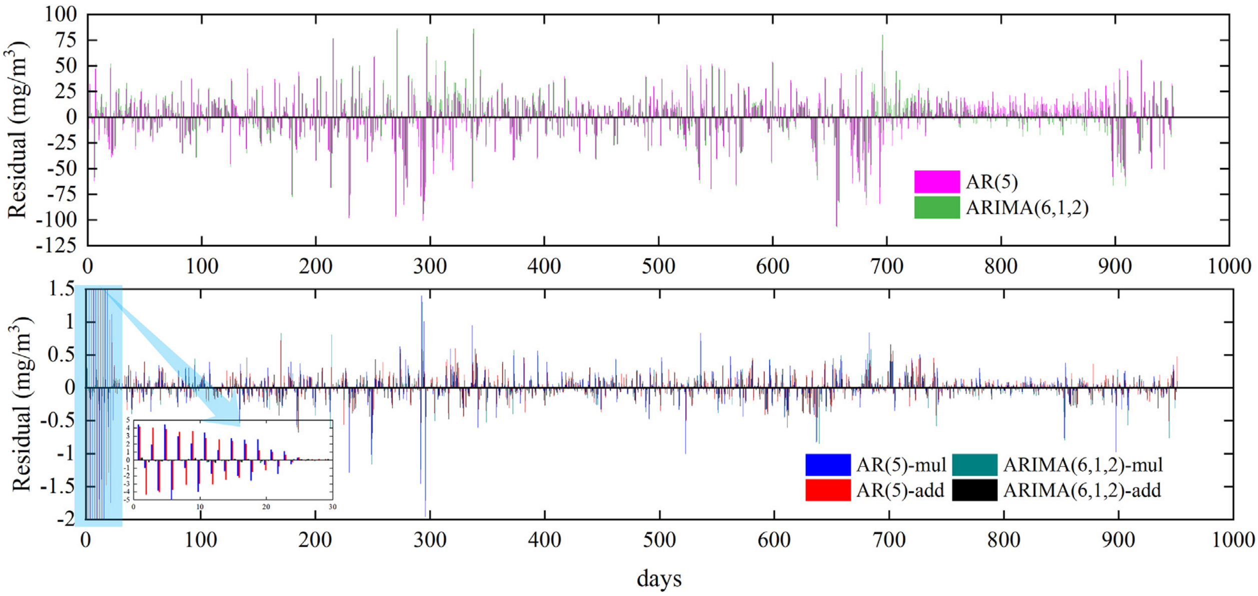


Figure 10. Residual variation diagram for the whole process of model test set.

the hybrid ARIMA model, and they are more evenly distributed throughout the test set predictions. However, for the hybrid ARIMA model using decomposed and reconstructed sequences, the residuals are larger mainly at the beginning of the prediction period (about the first 30 days), after which the prediction residuals increase slightly at the peak of $PM_{2.5}$ in 2016 (the serial number in the figure is about 300), except for which the prediction errors are extremely low for the whole process.

Also the blue and red color bars are longer compared to green and black, which is especially evident in the reduced plot for the first 30 days indicated by the arrow. This indicates that the ARIMA model outperforms the AR model in both multiplicative decomposition and additive decomposition.

4.5. Stability analysis and robustness check

The stability of the model is verified by plotting the scatter density of the full test set for the six models with good prediction performance and deriving the linear fit function as a comparison. Taylor plots are then plotted in order to further complete the robustness check.

Firstly, for the scatter density plot, the color of the points represents the density of the aggregated points, with red being the densest and blue being the sparsest, and the value of the right color bar indicates the normalized result of the point density. The advantage of normalized results is that data with different amounts of data can be represented by the same colorbar, for example, in this paper, the original data test set N = 996, due to seasonal decomposition, hybrid ARIMA model test set N = 951, but they share a colorbar. red solid line and black dashed line are linear fitted regression line and y = x reference line, respectively. The number of data N, the commonly used model accuracy evaluation metrics R^2 and RMSE, and our proposed comprehensive evaluation coefficient (CEC) are labeled in the upper left corner of the figure.

From Figure 11, all four hybrid ARIMA models predict better than the traditional ARIMA model according to the scatter plot distribution and the linear fitting function. From the scatter density case, it can be seen that the data are concentrated at 25 μg/m^3. By comparison, the ARIMA (6,1,2)-add model predicts the best results, where $R^2 = 0.99997$, RMSE = 0.18, CEC = 0.03783, and the fitting function is Prediction = 1.00004*Original-0.002.

Further analysis shows that only the slope of the ARIMA (6,1,2)-add model fit function is greater than 1. This indicates that the magnitude of most of the predicted values of this model exceeds the true value, but the slope of the fit function of all the remaining methods is less than 1. Combining the results in Figures 8 and 9: the traditional AR and ARIMA model fits poorly for data with large series volatility and abnormal extremes (very large or very small). There is a great improvement in the degree of fit of the hybrid ARIMA model, but most of the predicted data are still smaller than the true value. Only the ARIMA (6,1,2)-add model has the ability to make most of the predicted values exceed the true values, and it can achieve the up and down fluctuation of the predicted values with the true values as the center.

Taylor diagrams are then plotted to assess the robustness of the proposed model. Taylor diagrams are used to evaluate the accuracy of the model, and common Taylor diagram accuracy indicators are correlation coefficient R, standard deviation STD, and central root mean squared error (cRMSE) E' (Taylor, 2001). According to the literature, E' is calculated as.

$$E' = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[(X_i - \overline{X_i}) - (X'_i - \overline{X'_i}) \right]^2 \right\}^{1/2} \tag{26}$$

where X_i and X'_i are the true and predicted values of the time series respectively; n is the number of data in the test set.

Figure 12 presents the Taylor plot for the model comparison. The colored scatter in the Taylor plot represents the model, the blue radial line represents the correlation coefficient R, the gray solid line represents the standard deviation, and the green dashed line represents the central root mean square error (cRMSE). The advantage is the ability to use 3

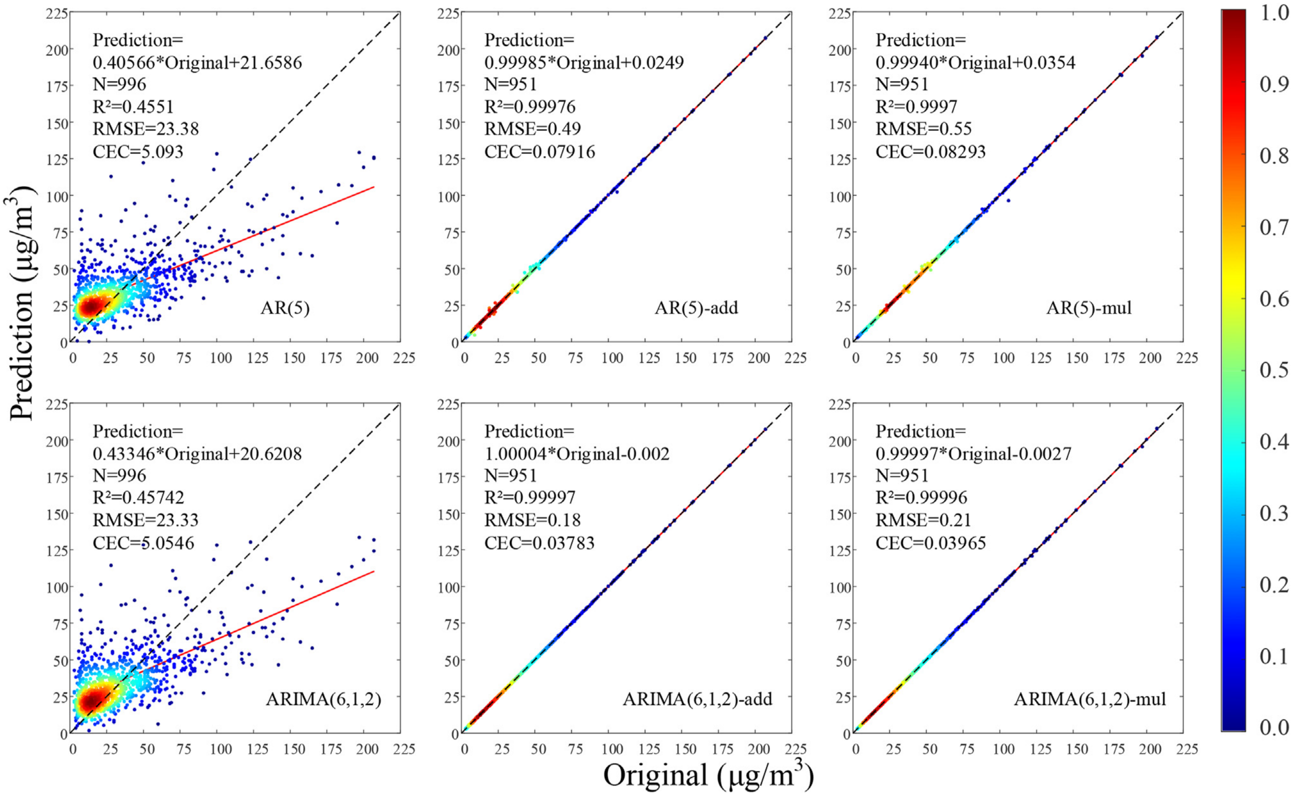


Figure 11. Correlation between the original and predicted $PM_{2.5}$ index via scatter density diagram.

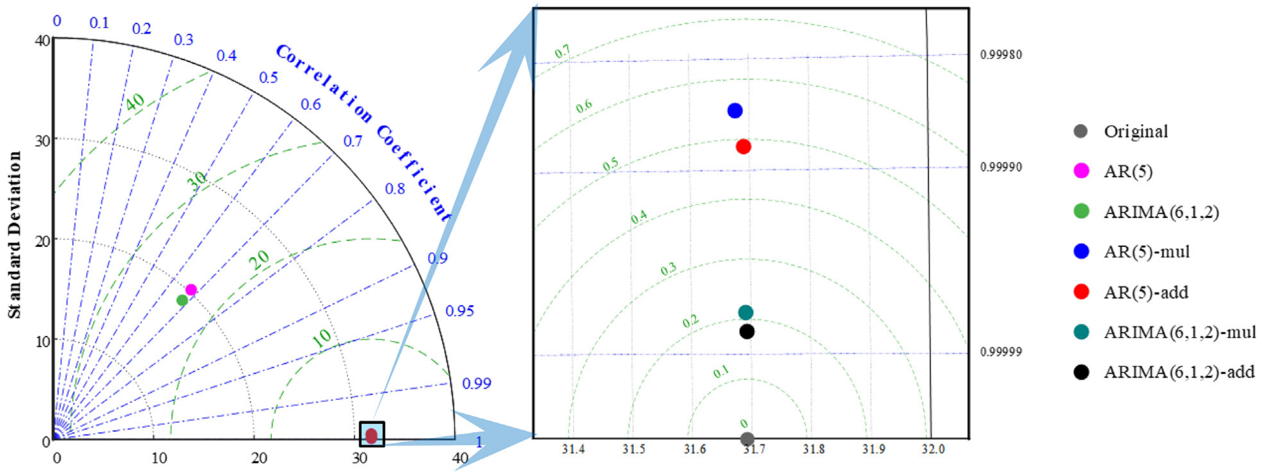


Figure 12. Taylor plots of original and hybrid ARIMA prediction models.

metrics to show the model accuracy. For the bottom right corner to zoom in, again from the figure it can be observed that the ARIMA (6,1,2)-add model has the best prediction performance and is closest to the Original point. Meanwhile, the results of the four models predicted by the hybrid ARIMA model are extremely close to the original data, which indicates that the prediction accuracy is significantly improved by using the seasonal decomposition, while the use of multiplicative or additive decomposition only slightly improves the prediction accuracy, which laterally reflects that the model has high stability.

4.6. Comparison with existing work

It will be very important to compare our improved model with the work of others, and in this paper we have chosen to compare and discuss the research paper by Shahriar et al. by presenting and comparing the accuracy of the models provided by both papers in a tabular format (Shahriar et al., 2021). The comparison results are shown in Table 11.

As mentioned in Table 11, In order to ensure the consistency of the data for the comparison, we have chosen the evaluation indicators RMSE, MAE and R^2 that are common in our work. The models with the best accuracy indicators for each city were bolded, namely the CatBoost model for Dhaka, Narayanganj and Gazipur and the ARIMA612-add model for Beijing.

Table 11. Comparison of model accuracy results against peer workers.

Sources	Regions	Models	RMSE	MAE	R^2
Work of Shahriar et al.	Dhaka	ARIMA-ANN	11.96000	6.78000	0.93000
		ARIMA-SVM	14.03000	8.51000	0.91000
		DT	12.27000	6.74000	0.88000
		CatBoost	**11.41000**	**5.82000**	**0.95000**
	Narayanganj	ARIMA-ANN	12.86000	7.64000	0.90000
		ARIMA-SVM	13.97000	8.31000	0.89000
		DT	13.07000	7.95000	0.89000
		CatBoost	**12.56000**	**6.97000**	**0.92000**
	Gazipur	ARIMA-ANN	12.34000	7.69000	0.91000
		ARIMA-SVM	12.68000	7.23000	0.89000
		DT	14.21000	7.97000	0.87000
		CatBoost	**12.07000**	**5.72000**	**0.94000**
Our works	Beijing	AR5	23.38220	16.71170	0.45510
		ARIMA612	23.33240	16.60210	0.45740
		AR5-mul	0.54774	0.18987	0.99970
		ARIMA612-mul	0.21101	0.12583	0.99996
		AR5-add	0.48789	0.19704	0.99976
		ARIMA612-add	**0.17950**	**0.13244**	**0.99997**

We found that Shahriar et al.'s CatBoost model does not differ much in prediction for different city datasets. The best accuracy index is the CatBoost model for Dhaka, with RMSE, MAE and R^2 values of 11.41000, 5.82000 and 0.95000 respectively, which is larger than that of our works. The ARIMA612-add model with integrated AIC and improved GS fixed-order methods and seasonal decomposition had RMSE, MAE and R^2 values of 0.17950, 0.13244, and 0.99997 respectively. comparison with the peer workers study, the algorithm and model we propose in this paper have even better performance.

5. Conclusion

In this study, a new hybrid ARIMA model is developed for predicting the daily average $PM_{2.5}$ concentration in Beijing. The data used in this study include eight years of historical time series from January 1, 2013 to January 31, 2022, of which 66% are classified as the training set and 34% are classified as the test set.

The hybrid ARIMA model proposed in this paper has four important innovations: (1) The ADF root test based on the annual $PM_{2.5}$ data is used to obtain the overall data smoothly, but some of the data are not smooth, so there is a need for the first-order difference. (2) The integrated AIC and improved GS methods are used to jointly determine the order, which avoids the bias caused by using AIC alone to determine the order because the data are not exactly normally distributed. The traditional AR and ARIMA models are also discussed for prediction. (3) The optimal AR and ARIMA model is selected by using the comprehensive evaluation coefficient (CEC), which is a comprehensive evaluation of a single evaluation index, to verify the accuracy and feasibility of the ranking method. (4) Finally, the original sequence is decomposed using seasonal decomposition and the entropy value of the decomposed sequence is obtained using RangeEn_A, which is reconstructed according to the size of the entropy value. Furthermore, the reconstructed sequence is used for prediction, in order to compare the original optimal AR and ARIMA model to evaluate the improvement of the hybrid ARIMA model in prediction performance.

The hybrid ARIMA model proposed in this paper has four important innovations: (1) The ADF root test based on the annual $PM_{2.5}$ data is used to obtain the overall data smoothly, but some of the data are not smooth, so there is a need for the first-order difference. (2) The integrated AIC and improved GS methods are used to jointly determine the order, which avoids the bias caused by using AIC alone to determine the order because the data are not exactly normally distributed. The traditional AR and ARIMA models are also discussed for prediction. (3) The optimal AR and ARIMA model is selected by using the comprehensive evaluation

coefficient (CEC), which is a comprehensive evaluation of a single evaluation index, to verify the accuracy and feasibility of the ranking method. (4) Finally, the original sequence is decomposed using seasonal decomposition and the entropy value of the decomposed sequence is obtained using RangeEn_A, and the reconstruction is carried out according to the entropy value magnitude. And the reconstructed sequence is used for prediction, and the original optimal AR and ARIMA model are compared to evaluate the improvement of the hybrid ARIMA model in prediction performance.

Compared with the ARIMA model, the proposed hybrid ARIMA (including ARIMA-add and ARIMA-mul) model has better prediction performance. Among them, the ARIMA-add model is the most advanced, which has the ability to make most of the predicted values exceed the true values and can achieve up and down fluctuations of the predicted values centered on the true values. In addition, we also conducted a stability analysis and robustness check, and the results show that the prediction accuracy is substantially improved by using the seasonal decomposition.

In summary, the validation results show that the new hybrid ARIMA model has good prediction performance. 99.23% improvement in RMSE, 99.20% improvement in MAE, 118.61% improvement in R^2, 99.28% improvement in TIC, 98.71% improvement in NMAE, 99.97% improvement in NMSE, and 43.13% improvement in OPC, The CEC index is improved by 99.25%. Both in prediction accuracy, directional prediction accuracy and model fit are substantially improved compared with the traditional ARIMA model. The method can be applied as a convincing analysis and prediction tool in practical fields. The derived $PM_{2.5}$ is beneficial to the formulation of relevant regulatory policies to reduce regional air pollution levels and provide practical protection for people's life and health. However, $PM_{2.5}$ largely depends on many factors, such as vehicle emissions, building construction and industrial activity intensity. Therefore, there are still some research questions that need further study. The model can also attach some external influences, such as sulfur dioxide index, nitrogen dioxide index and industrial exhaust emissions in the area, to improve the prediction performance.

Declarations

Author contribution statement

Lingxiao Zhao: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Zhiyang Li: Data curation; Formal analysis; Performed the experiments.

Leilei Qu: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data.

Funding statement

Lingxiao Zhao was supported by National College Students Innovation and Entrepreneurship Training Program [202110158002], Liaoning College Student Innovation and Entrepreneurship Training Program Fund [S202210158006].

Dr. Leilei Qu was supported by Doctoral Start-up Foundation of Liaoning Province [2020-BS-216], Foundation of Liaoning Province Education Administration [LJKMZ20221115].

Data availability statement

Data will be made available on request.

Declaration of interests statement

The authors declare no competing interests.

Additional information

No additional information is available for this paper.

Acknowledgements

The authors also express their deep gratitude to the editors and reviewers for their careful reading, suggestions and comments on the manuscript.

References

Abbaszadeh, M., Soltani-Mohammadi, S., Ahmed, A.N., 2022. Optimization of support vector machine parameters in modeling of Iju deposit mineralization and alteration zones using particle swarm optimization algorithm and grid search method. Comput. Geosci., 105140

Akhter, M.F., Hassan, D., Abbas, S., 2020. Predictive ARIMA Model for coronal index solar cyclic data. Astron. Comput. 32, 100403.

Aladağ, E., 2021. Forecasting of particulate matter with a hybrid ARIMA model based on wavelet transformation and seasonal adjustment. Urban Clim. 39, 100930.

Arora, S., Keshari, A.K., 2021. ANFIS-ARIMA modelling for scheming re-aeration of hydrologically altered rivers. J. Hydrol. 601, 126635.

Belavadi, S.V., Rajagopal, S., R, R., Mohan, R., 2020. Air quality forecasting using LSTM RNN and wireless sensor networks. Procedia Comput. Sci. 170, 241–248.

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G., Di Carlo, P., 2017. Recursive neural network model for analysis and forecast of PM10 and PM2.5. Atmos. Pollut. Res. 8 (4), 652–659.

Bisaglia, L., Procidano, I., 2002. On the power of the Augmented Dickey–Fuller test against fractional alternatives using bootstrap. Econ. Lett. 77 (3), 343–347.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control. John Wiley & Sons.

Çankaya, M.N., Korbel, J., 2018. Least informative distributions in maximum q-log-likelihood estimation. Phys. Stat. Mech. Appl. 509, 140–150.

Catalano, M., Galatioto, F., Bell, M., Namdeo, A., Bergantino, A.S., 2016. Improving the prediction of air pollution peak episodes generated by urban transport networks. Environ. Sci. Pol. 60, 69–83.

Chang, Z., Yuan, W., Huang, K., 2022. Remaining useful life prediction for rolling bearings using multi-layer grid search and LSTM. Comput. Electr. Eng. 101, 108083.

Chen, F., Deng, Z., Deng, Y., Qiao, Z., Lan, L., Meng, Q., Luo, B., Zhang, W., Ji, K., Qiao, X., Fan, Z., Zhang, M., Cui, Y., Zhao, X., Li, X., 2017. Attributable risk of ambient PM10 on daily mortality and years of life lost in Chengdu, China. Sci. Total Environ. 581–582, 426–433.

Cheng, Y., Zhang, H., Liu, Z., Chen, L., Wang, P., 2019. Hybrid algorithm for short-term forecasting of PM2.5 in China. Atmos. Environ. 200, 264–279.

Cobbold, A.T., Crane, M.A., Knibbs, L.D., Hanigan, I.C., Greaves, S.P., Rissel, C.E., 2022. Perceptions of air quality and concern for health in relation to long-term air pollution exposure, bushfires, and COVID-19 lockdown: a before-and-after study. J. Clim. Change Health, 100137.

Dai, H., Ma, D., Zhu, R., Sun, B., He, J., 2019. Impact of control measures on nitrogen oxides, sulfur dioxide and particulate matter emissions from coal-fired power plants in Anhui Province, China. Atmosphere 10 (1).

Dong, J., Wang, Y., Wang, L., Zhao, W., Huang, C., 2022. Assessment of PM2.5 exposure risk towards SDG indicator 11.6.2 – a case study in Beijing. Sustain. Cities Soc. 82, 103864.

Du, P., Wang, J., Hao, Y., Niu, T., Yang, W., 2020. A novel hybrid model based on multi-objective Harris hawks optimization algorithm for daily PM2.5 and PM10 forecasting. Appl. Soft Comput. 96, 106620.

Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., Wang, J., 2015. Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. 107, 118–128.

Guo, J.-X., Zeng, Y., Zhu, K., Tan, X., 2021. Vehicle mix evaluation in Beijing's passenger-car sector: from air pollution control perspective. Sci. Total Environ. 785, 147264.

He, Q., Gao, K., Zhang, L., Song, Y., Zhang, M., 2021. Satellite-derived 1-km estimates and long-term trends of PM2.5 concentrations in China from 2000 to 2018. Environ. Int. 156, 106726.

Huang, W., Cai, L., Dang, H., Jiao, Z., Fan, H., Cheng, F., 2019. Review on formation mechanism analysis method and control strategy of urban haze in China. Chin. J. Chem. Eng. 27 (7), 1572–1577.

Kärner, O., 2009. ARIMA representation for daily solar irradiance and surface air temperature time series. J. Atmos. Sol. Terr. Phys. 71 (8), 841–847.

Kębłowski, P., Welfe, A., 2004. The ADF–KPSS test of the joint confirmation hypothesis of unit autoregressive root. Econ. Lett. 85 (2), 257–263.

Kong, Y., Sheng, L., Li, Y., Zhang, W., Zhou, Y., Wang, W., Zhao, Y., 2021. Improving PM2.5 forecast during haze episodes over China based on a coupled 4D-LETKF and WRF-Chem system. Atmos. Res. 249, 105366.

Matsumoto, C., Yanagihara, H., Wakaki, H., 2011. Improvement of the quality of the chi-square approximation for the ADF test on a covariance matrix with a linear structure. J. Stat. Plann. Inference 141 (4), 1535–1542.

Mills, T.C., 2019. Chapter 3 - ARMA models for stationary time series. In: Mills, T.C. (Ed.), Applied Time Series Analysis. Academic Press, pp. 31–56.

Ning, Y., Kazemi, H., Tahmasebi, P., 2022. A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet. Comput. Geosci. 164, 105126.

Omidvarnia, A., Mesbah, M., Pedersen, M., Jackson, G., 2018. Range entropy: a bridge between signal complexity and self-similarity. Entropy 20 (12), 962.

Shahriar, S.A., Kayes, I., Hasan, K., Hasan, M., Islam, R., Awang, N.R., Hamzah, Z., Rak, A.E., Salam, M.A., 2021. Potential of ARIMA-ANN, ARIMA-SVM, DT and CatBoost for atmospheric PM2.5 forecasting in Bangladesh. Atmosphere 12 (1), 100.

Sharma, S., Zhang, M., Anshika, Gao, J., Zhang, H., Kota, S.H., 2020. Effect of restricted emissions during COVID-19 on air quality in India. Sci. Total Environ. 728, 138878.

Snipes, M., Taylor, D.C., 2014. Model selection and Akaike Information Criteria: an example from wine ratings and prices. Wine Econ. Pol. 3 (1), 3–9.

Suárez Sánchez, A., García Nieto, P.J., Riesgo Fernández, P., del Coz Díaz, J.J., Iglesias-Rodríguez, F.J., 2011. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). Math. Comput. Model. 54 (5), 1453–1466.

Sun, J., 2021. Forecasting COVID-19 pandemic in Alberta, Canada using modified ARIMA models. Comp. Method. Prog. Biomed. Update 1, 100029.

Taddy, M., 2019. Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions. McGraw-Hill Education.

Tang, C.-S., Wu, T.-Y., Chuang, K.-J., Chang, T.-Y., Chuang, H.-C., Lung, S.-C.C., Chang, L.-T., 2019. Impacts of in-cabin exposure to size-fractionated particulate matters and carbon monoxide on changes in heart rate variability for healthy public transit commuters. Atmosphere 10 (7).

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atmos. 106 (D7), 7183–7192.

Theerthagiri, P., 2022. Mobility prediction for random walk mobility model using ARIMA in mobile ad hoc networks. J. Supercomput. 78 (14), 16453–16484.

Tian, X., Zhang, Y., Lin, Z., 2022. Predicting non-uniform indoor air quality distribution by using pulsating air supply and SVM model. Build. Environ. 219, 109171.

Vohra, K., Vodonos, A., Schwartz, J., Marais, E.A., Sulprizio, M.P., Mickley, L.J., 2021. Global mortality from outdoor fine particle pollution generated by fossil fuel combustion: results from GEOS-Chem. Environ. Res. 195, 110754.

Wang, D., Wei, S., Luo, H., Yue, C., Grunder, O., 2017a. A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Sci. Total Environ. 580, 719–733.

Wang, J., Yang, Y., Jiang, X., Wang, D., Zhong, J., Wang, Y., 2022. Observational study of the PM2.5 and O3 superposition-composite pollution event during spring 2020 in Beijing associated with the water vapor conveyor belt in the northern hemisphere. Atmos. Environ. 272, 118966.

Wang, P., Zhang, H., Qin, Z., Zhang, G., 2017b. A novel hybrid-Garch model based on ARIMA and SVM for PM2.5 concentrations forecasting. Atmos. Pollut. Res. 8 (5), 850–860.

Wang, T., Jiang, F., Deng, J., Shen, Y., Fu, Q., Wang, Q., Fu, Y., Xu, J., Zhang, D., 2012. Urban air quality and regional haze weather forecast for Yangtze River Delta region. Atmos. Environ. 58, 70–83.

Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., Chi, T., 2019. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. Sci. Total Environ. 654, 1091–1099.

Wen, X., Chen, W., Chen, B., Yang, C., Tu, G., Cheng, T., 2020. Does the prohibition on open burning of straw mitigate air pollution? An empirical study in Jilin Province of China in the post-harvest season. J. Environ. Manag. 264, 110451.

World Health, O., 2016. Ambient Air Pollution: a Global Assessment of Exposure and burden of Disease. World Health Organization, Geneva.

World Health Organization, 2002. Regional office for, E. In: Health Impact Assessment of Air Pollution in the Eight Major Italian Cities. WHO Regional Office for Europe, Copenhagen.

World Health Organization, 2003. Exposure Assessment in Studies on the Chronic Effects of Long-Term Exposure to Air Pollution : Report on a WHO/HEI Workshop, Bonn. Regional Office for, E. WHO Regional Office for Europe, Copenhagen. Germany 4-5 February 2002.

Yang, J., Tang, Y., Han, S., Liu, J., Yang, X., Hao, J., 2021. Evaluation and improvement study of the Planetary Boundary-Layer schemes during a high PM2.5 episode in a core city of BTH region, China. Sci. Total Environ. 765, 142756.

Yang, L., Gao, X., Li, Z., Jia, D., 2022a. Quantitative effects of air pollution on regional daily global and diffuse solar radiation under clear sky conditions. Energy Rep. 8, 1935–1948.

Yang, X., Xiao, D., Fan, L., Li, F., Wang, W., Bai, H., Tang, J., 2022b. Spatiotemporal estimates of daily PM2.5 concentrations based on 1-km resolution MAIAC AOD in the Beijing–Tianjin–Hebei, China. Environ. Challenge. 8, 100548.

Zafra, C., Ángel, Y., Torres, E., 2017. ARIMA analysis of the effect of land surface coverage on PM10 concentrations in a high-altitude megacity. Atmos. Pollut. Res. 8 (4), 660–668.

Zeng, S., Liu, H., Liu, Z., Kaufmann, G., Zeng, Q., Chen, B., 2019. Seasonal and diurnal variations in DIC, NO3− and TOC concentrations in spring-pond ecosystems under different land-uses at the Shawan Karst Test Site, SW China: carbon limitation of aquatic photosynthesis. J. Hydrol. 574, 811–821.

Zhan, J., Wu, C., Ma, X., Yang, C., Miao, Q., Wang, S., 2022. Abnormal vibration detection of wind turbine based on temporal convolution network and multivariate coefficient of variation. Mech. Syst. Signal Process. 174, 109082.

Zhang, L., Lin, J., Qiu, R., Hu, X., Zhang, H., Chen, Q., Tan, H., Lin, D., Wang, J., 2018. Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model. Ecol. Indicat. 95, 702–710.

Zhang, M.-H., Cheng, Q.-S., 2004. Determine the number of components in a mixture model by the extended KS test. Pattern Recogn. Lett. 25 (2), 211–216.

Zhang, X., Chen, L., Yuan, R., 2020. Effect of natural and anthropic factors on the spatiotemporal pattern of haze pollution control of China. J. Clean. Prod. 251, 119531.

Zhao, J., Deng, F., Cai, Y., Chen, J., 2019. Long short-term memory - fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. Chemosphere 220, 486–492.

Zhou, W., Wu, X., Ding, S., Ji, X., Pan, W., 2021. Predictions and mitigation strategies of PM2.5 concentration in the Yangtze River Delta of China based on a novel nonlinear seasonal grey model. Environ. Pollut. 276, 116614.

Zhou, Y., Chen, J., Yu, Z., Zhou, J., Zhang, G., 2022. Short-term building occupancy prediction based on deep forest with multi-order transition probability. Energy Build. 255, 111684.