Predicting Pressure from Environmental Factors in London based on ARIMAX Model

Zhizhi Li[1], Qi Zhang[2,*] and Zhengxi Zhou[3]

[1]Kunming NO.8 High School, Kunming, 650032, China

[2]School of Mathematics, Nanjing Xiaozhuang University, Nanjin, 211171, China

[3]Suzhou foreign language school, Suzhou, 215011, China

*Corresponding author: 20131007@njxzc.edu.cn

Abstract. Pressure is an important index in atmospheric science as it influences various aspects of both the natural environment and everyday life. Although the significance of pressure is widely recognized, accurately measuring and recording pressure data remains a crucial issue. This is due to the fact that weather prediction involves fitting a large number of nonlinear factors, which poses significant challenges in establishing precise predicting models. Therefore, the current focus of most research on pressure lies in the innovation and improvement of algorithms for pressure sensors and other related devices. This paper makes a statistical analysis on the raw climate data of London over a period of 13 years starting from 2008, applying an ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables) dynamic regression model to fit 4 exogenous variables to the pressure data as linear factors. The fitting results showed that the variation in pressure is associated with multiple environmental factors. Furthermore, it is observed that the pressure in London is projected to be slightly higher than normal one in the upcoming months. Due to the numerous parameters in the ARIMAX model, overfitting and underfitting phenomena are prone to occur, therefore, for parameter adjustments, further processing is required.

Keywords: Pressure; environmental factors; ARIMAX model.

1. Introduction

Air pressure refers to the vertical force exerted by the air on the Earth's surface or other objects, typically measured in units of pressure per unit area, such as pascals. As an important indicator in Atmospheric Science, air pressure plays a critical role in a large number of fields such as Meteorology, Environmental Science, climate research, and weather forecasting, et al. [1]. It is of significant importance for understanding and predicting weather patterns [2], explaining climate change, and assessing environmental impacts [3].

The research on pressure has received widespread attention in various fields over the past few decades, the establishment of weather stations and satellite observation networks has provided more precise data sources for pressure observations. In fact, pressure plays an important role in environmental research and monitoring [4], it can be used to assess and monitor air pollution, atmospheric quality [5] and the impact of climate change on the environment [3]. The changes in atmospheric pressure also can affect the stability of the atmosphere and wind speed [6], it is a direct impact on air quality and the dispersion of pollutants [7]. Therefore, as an important daily indicator, pressure holds great significance for weather forecasting, climate research, and environmental monitoring. By comprehensively understanding and studying atmospheric pressure, scientists can gain a better understanding of the operational mechanisms of the Earth atmospheric system [8] and provide scientific basis and decision support for addressing climate change and environmental challenges.

There are a number of factors that affect the change of air pressure. Climate change can lead to an increase in the frequency of formation cyclone which will result in changes in atmospheric pressure [9], tropical cyclones sometimes even have destructive impacts on the natural environment. Also, during periods of atmospheric circulation transition, there is a significant exchange of heat between high and low latitudes in order to achieve thermal equilibrium which will cause mutative atmospheric

pressure [10]. Some scholars explore the impact of environmental factors on air pressure by studying climate change patterns and variations in climate [11]. Therefore, this paper mainly studies 4 exogenous factors of environmental elements in about 22 years (Cloud cover, Sunshine, mean Temperature, Precipitation), and choose a suitable statistical model to study these factors and predict pressure.

International and domestic Scholars have established numerous models and algorithms for researching and predicting pressure, Liu et al. analyze the relationship between the standard air pressure value and the errors of the pressure sensor, deriving an error fluctuation description formula, and she designed an RBF model for fitting error coefficients which effectively improves the measurement accuracy of the pressure sensors used in high-altitude meteorological detection [12]. But deep learning models are not lightweight enough, which leads to a decrease in the calibration precision of the pressure sensors. Gen et al. evaluated the applicability of ERA5 reanalysis data and found a good correlation between the reanalysis data with the observed data of sea-level pressure and wind speed and the correlation changes with climate variations, but under extreme weather conditions, the applicability of reanalysis data for pressure and wind speed is poor, making it difficult to evaluate their correlation [13]. For the issue of significant periodic errors in various meteorological parameters of the Global Pressure and Temperature (GPT2w) model, Xu et al. utilized a multi-order Fourier function to establish a new improved GPT2w model which resulted in a significant improvement in data accuracy for the model in the Yangtze River Delta region [14].

In summary, after studying the data features and the related literatures, the paper will use the ARIMAX model to analyse a climate dataset consisting of a sample size of 8402 and predict the mean air pressure from 4 environmental factors in London for the day of the end of 2023.

2. Methods

2.1. Data Sources

The data set is collected from the Kaggle website and is completed by Emmanuel F. Werr in 2023, which is created by reconciling measurements from requests of individual weather attributes provided by the European Climate Assessment (ECA) from 1998 to 2020 in London, UK.

2.2. Variable Selection

The paper just counts the data in London and the data contains 4 different variables (Cloud cover, Sunshine, mean Temperature, Precipitation), as Table 1 and 2 show.

Table 1. Weather statistics in London

date	cloud cover	sunshine	mean temp	precipitation	pressure
19980101	6	3.5	6.2	12.4	100190
19980102	5	0.9	8.5	7	98590
19980103	6	2.3	9	5.8	99120
19980104	6	0.1	6.2	12	99170
19980105	5	2.6	5	6.6	99480
...	...	...	...	...	...
20201227	1	0.9	7.5	2	98000
20201228	7	3.7	1.1	0.2	97370
20201229	7	0	2.6	0	98830
20201230	6	0.4	2.7	0	100200
20201231	7	1.3	-0.8	0	100500

Table 2. Logogram of dependent variable and exogenous variables

Variables	Logogram	Size
cloud cover	x_1	8402
sunshine	x_2	8402
mean temp	x_3	8402
precipitation	x_4	8402
pressure	y	8402

2.3. Research Protocol

This paper uses the ARIMAX time series model. The mean air pressure of London in the past 22 years is dependent variable (Y), using 4 exogenous factors as independent variables (X). Next, this paper uses the model established by SPSS to predict the mean pressure of London to the end of 2023.

2.3.1 ARIMAX model

ARIMAX is a time series forecasting model that combines the ARIMAX model with the influence of exogenous variables. ARIMAX model is a common time series modeling method for capturing the autocorrelation properties and moving average properties of the time series. It can be seen as ARIMA $(\mathrm{p},\mathrm{d},\mathrm{q})(\mathrm{P},\mathrm{D},\mathrm{Q},\mathrm{S})$. Where $\mathrm{p},\mathrm{d},\mathrm{q}$ are the non-seasonal order-numbers, respectively denoting autoregressive order, differencing order, moving average order and $\mathrm{P},\mathrm{D},\mathrm{Q},\mathrm{S}$ are the seasonal order-numbers, respectively denoting seasonal autoregressive order, seasonal differencing order, seasonal moving average order and seasonal periodicity of statistics. Seeing this prediction is based on a time series dataset with an annual time unit, this paper does not consider seasonal factors.

The ARIMAX model is a variant of the autoregressive integrated moving average model (ARIMA) model. By introducing exogenous variables, the ARIMAX model allows for the consideration of other factors that impact the time series, which is required to difference the unstable time series in the ARIMA model to establish an ARMA model. And the ARMA model is a comprehensive model that combines the autoregressive (AR) model and Moving Average (MA) model.

2.3.2 AR model algorithm

The AR model is an p-order autoregressive model, utilizing the value of x_t and x_{t-1} to predict that of x_t for the form of an autoregressive model is to describe the relationship between a variable and its past values. Due to the linearity of the values of x, and the form of a p-order autoregressive model can be represented as

$$x_i = c + \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t \tag{1}$$

Where c is constant term and ϵ_t is a random error term with a mean value of zero and a standard deviation of σ. After introducing the delay operator B and centralizing the AR model, the polynomial of p-order auto-regressive coefficients can be transformed as

$$\phi(\mathrm{B}) = 1 - \phi_1 \mathrm{B} - \phi_2 \mathrm{B}^2 - \cdots - \phi_\mathrm{p} \mathrm{B}^\mathrm{p} \tag{2}$$

2.3.3 MA model algorithm

The MA model is a q-order moving average model, utilizing the mean value of x_t and x_{t-1} to predict that of x_t.

$$x_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{i-1} \tag{3}$$

Where μ is series mean, $\theta_1, \ldots, \theta_q$ are parameters and $\epsilon_t, \ldots, \epsilon_{t-q}$ are white noise. After introducing the delay operator B and centralizing the MA model, the polynomial of q-order moving average coefficients can be transformed as

$$\psi(\mathrm{B}) = 1 - \theta_1 \mathrm{B} - \theta_2 \mathrm{B}^2 - \cdots - \theta_\mathrm{q} \mathrm{B}^\mathrm{q} \tag{4}$$

2.3.4 ARIMAX model algorithm

The ARIMAX model is an ARIMA model with input variables. When establishing an ARIMAX model, the first consideration is the stationarity of the sequence. If the sequence is non-stationary, difference is applied to the independent variable $\{y_t\}$ and $\{x_{it}\}$ sequence to transform it into a stationary sequence $\{\nabla_{yt}^{\iota}\}$ and $\{\nabla_{xit}^{\iota}\}$.

$$\nabla_{yt}^{\iota} = \mu + \sum_{i=1}^{t} \frac{\psi_i(B)}{\phi_i(B)} B^{\iota i} \nabla_{xit}^{\iota} + \epsilon_t, \text{ where } \epsilon_t = \frac{\psi_i(B)}{\phi_i(B)} a_t \tag{5}$$

where B is retardation factor, $\phi_i(B)$ is the polynomial of autoregressive coefficient for the i-th input sequence, $\psi_i(B)$ is the polynomial of moving average coefficient for the i-th input sequence, ι_i is the delay order for the i-th input variable and ϵ_k is the regression residuals sequence. For $\{y_t\}$ and $\{x_{it}\}$ become stationary sequences after l-order difference and stationary sequences have linear property, therefore the residual sequence is also a stationary sequence. It can be seen as

$$\epsilon_t = y_t - (\mu + \sum_{i=1}^{k} \frac{\psi_i(B)}{\phi_i(B)} B^{\iota i} x_{it}) \tag{6}$$

And the model structure of ARIMAX can be expressed as

$$\begin{cases} \nabla_{yt}^{\iota} = \mu + \sum_{i=1}^{t} \frac{\psi_i(B)}{\phi_i(B)} B^{\iota i} \nabla_{xit}^{\iota} + \epsilon_t \\ \epsilon_t = \frac{\psi_i(B)}{\phi_i(B)} a_t \end{cases} \tag{7}$$

Where $\phi(B)$ is polynomial of auto-regressive coefficient of residual sequences, $\psi(B)$ is polynomial of moving average coefficient of residual sequences and a_t is a white noise sequence with a mean value of zero.

3. Results and Discussion

3.1. Model Establishment

Conduct stationarity test on the dependent variable sequence $\{y_t\}$ and independent variable sequences $\{x_{it}\}$, and plot a time series graph of the input sequence. A white noise image that fluctuates around the value of 101,000 with consistent magnitude and direction of the fluctuations which is a stationary sequence. From Fig 1, it can be seen as a white noise sequence with a certain trend and that the input is a stationary sequence, therefore it is no need to operate difference, and directly consider the autoregressive order and moving average order.

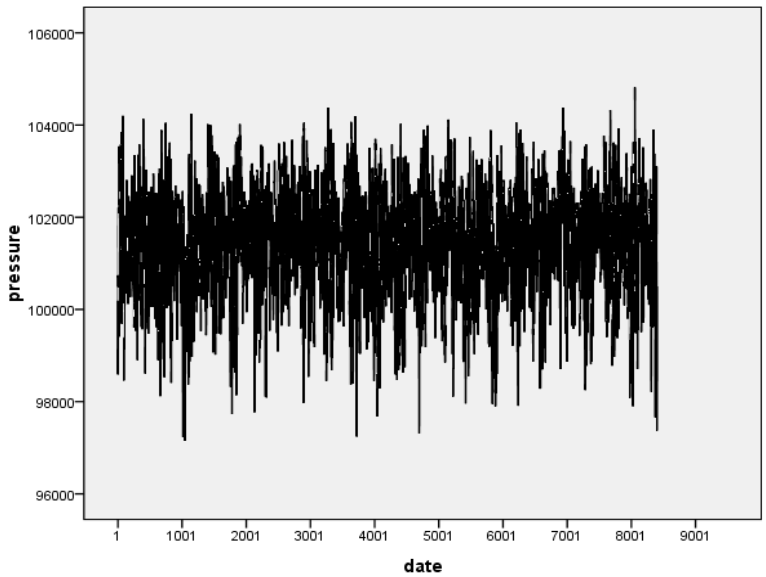

Fig 1. The original time sequence chart

From Fig 2 and Fig 3, it can be seen that the autocorrelation function of the input sequence has a tail and gradually approaches to 0. Additionally, the partial autocorrelation function has a significant 3-order truncation. And both images have a definite trend, with relatively small fluctuations in mean and variance, therefore, the original time series can be judged as a stationary sequence, where p is equal to 1 and q is equal to 3, and establish ARIMA (1,0,3) model.

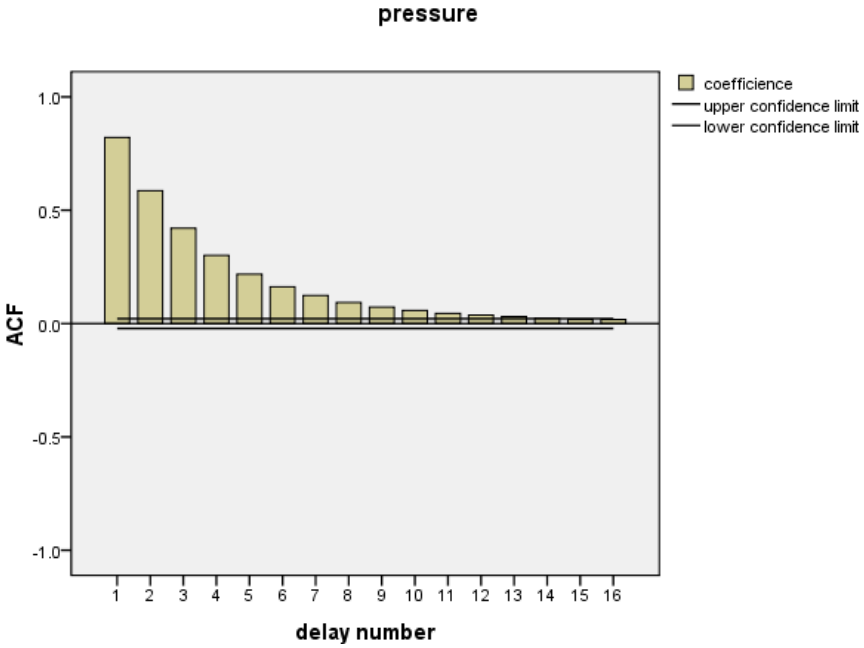

Fig 2. Original time sequence autocorrelation function chart

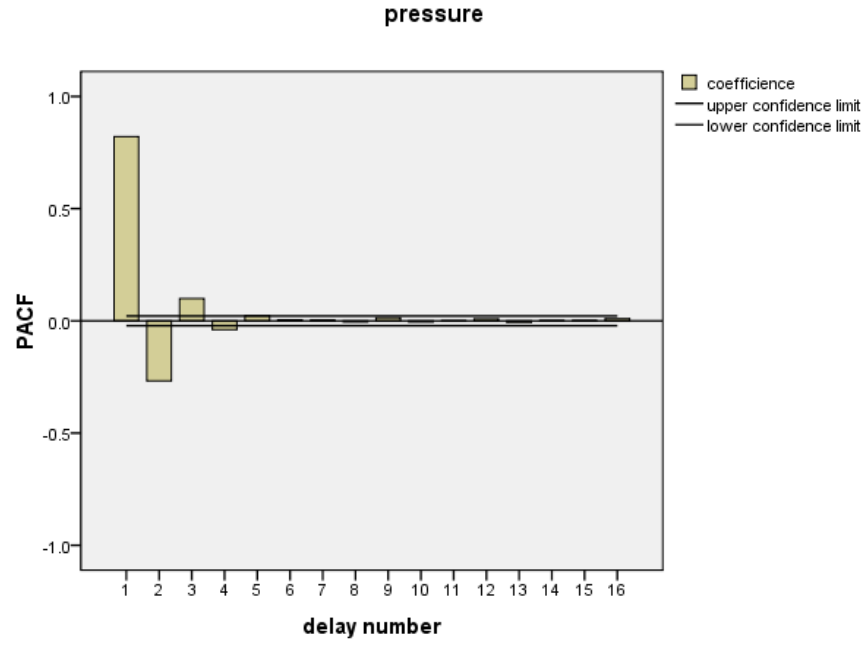

Fig 3. Original time sequence partial autocorrelation function chart

3.2. Model Checking

When utilizing the "Expert Modeler" function in SPSS for modeling, it is a request to select the original time sequence as the dependent variable and add the four exogenous variables as independent variables. Based on whether outlier detection is operated, Two sets of model statistics data can be obtained, as shown in Table 3 and Table 4, Table 3 shows the data when outliers are not detected, it can be seen that the R^2 is 0.762 which is a little lower than 0.771 of Table 4, in terms of significance, both values are over 0.05, therefore statistical differences exist.

Table 3. ARIMAX prediction accuracy summary

model statistics							
model	number of predictive variables	model fitting statistics		Ljung-Box Q(18)			outlier number
		stable R^2	R^2	statistics	DF	Sig	
pressure model1	4	0.762	0.762	23.840	15	0.068	0

Table 4. ARIMAX prediction accuracy summary (outlier number checking)

model statistics							
model	Number of predictive variables	model fitting statistics		Ljung-Box Q(18)			outlier number
		Stable R^2	R^2	statistics	DF	Sig.	
Pressure model1	4	0.771	0.771	21.526	15	0.121	12

3.3. Results Analysis

After the model establishment and model checking, utilizing ARIMA(1,0,3) to fit the input variables and exogenous variables, and making prediction. Next, it is a need to test the residual autocorrelation and residual partial autocorrelation of the fitting function. When a significant number of data points for both residual autocorrelation and partial autocorrelation are outside the 95% confidence interval, it is necessary to reanalyze the p-order autoregressive and the q-order moving average. From Fig 4, the ARIMA (1, 0, 3) model fitting the residual autocorrelation function (ACF) and residual partial autocorrelation function (PACF) are almost entirely within the 95% confidence interval. It can be seen that the residual autocorrelation and partial autocorrelation function are truncated, therefore, the paper can conclude that the residual ACF sequence and the residual PACF sequence are both white noise sequences and the correlation of the sequence has been fully fitted.

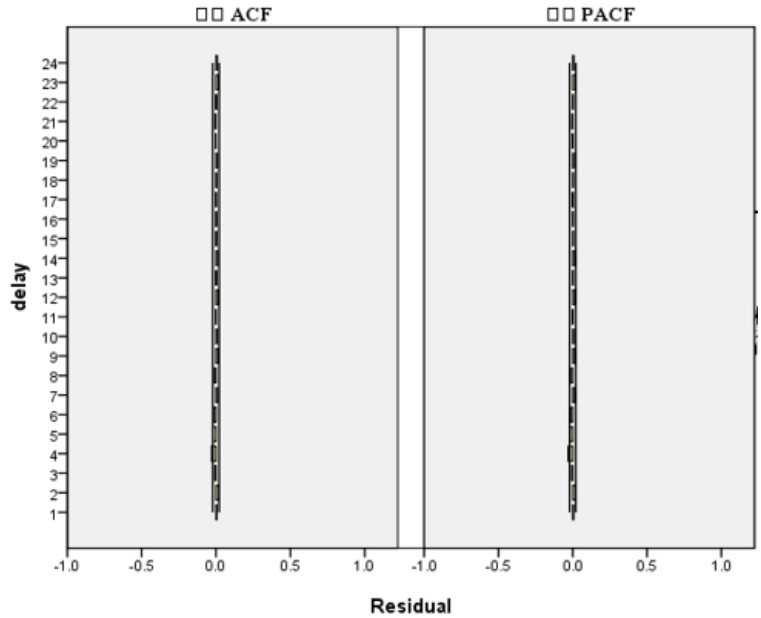

Fig 4. Residual ACF and residual PACF

The ARIMA model is a comprehensive model composed of autoregressive (AR) model and moving average (MA) model. From Table 5, it can be observed that the significance level for the 1-order delay autoregressive (AR) model and the 3-order delay moving average (MA) model are both lower than 0.05, it can be judged that the chosen p-order autoregressive and q-order moving average are reasonable. For the model parameters of the independent variables, there is obvious difference in the significance testing of the four exogenous variables. This indicates that the model parameters fit well, suggesting a relatively good model fitting. Based on the values of the model parameter delay and estimation, as well as extensive analysis and calculations, the fitting function of the model can be presented as

$$P(t) = 0.71P(t-1) + 0.033\epsilon(t) + 170.432F(cloud_{cover}) + 13.882F(sunshine) + 267.549F(mean\_temp) + 83.154F(precipitation) \quad (8)$$

Table 5. Model parameters

Variable	Form	Delay Number	Estimation	SE	t	Sig
pressure	constant		102166.638	82.653	1236.086	0.000
	constant	Delay 1	.711	0.010	71.552	0.000
	AR	Delay 1	-.326	0.013	-26.073	0.000
	MA	Delay 3	033	0.011	3.062	0.002
SMEAN (cloud cover1)	-	Delay 0	-45.676	3.860	-11.834	0.000
	numerator	Delay 1	543	0.062	8.775	0.000
	denominator	Delay 2	-.268	0.061	-4.353	0.000
SMEAN (sunshine1)	-	Delay 0	13.882	1.874	7.408	0.000
SMEAN (mean temp1)	numerator	Delay 0	66.940	3.001	-22.303	0.000
	numerator	Delay 1	-65.817	3.058	-21.523	0.000
	-	Delay 1	1.142	0.045	25.625	0.000
	denominator	Delay 2	-.246	0.043	-5.669	0.000
SMEAN (precipitation1)	-	Delay 0	-40.080	1.591	-25.199	0.000
	numerator	Delay 1	1.220	0.029	42.414	0.000
	denominator	Delay 2	-.482	0.028	-17.164	0.000

4. Conclusion

This paper uses the ARIMAX dynamic regression model. This paper analyzes the influence of 4 different exogenous variables (Cloud cover, mean Temperature, Sunshine, Precipitation) on the input variable. By analyzing the original time sequence diagram, it can be found that the sequence diagram is a white noise image, which fluctuates stably around 101,000 pascals. Therefore, the original time sequence can be considered as a stable sequence without difference processing. Then, by judge the tailing and truncation characteristics of the autocorrelation function and partial autocorrelation function of the original time sequence, the model ARIMA (1,0,3) is established. The autoregressive order is 1, for a certain moment in the future is related to the value of the previous moment, and the observed value of the past moment is used to predict the future. The difference order is 0, for the original data set is stationary, so the ARMA model can be constructed by using stationary time series data. The moving average order is 3, for the 3-order moving average factor is added to predict future data. This paper uses the "expert simulator" to analyze the outliers and make statistics to output the model fitting table. This paper finds that the R^2 reaches 0.771, which basically meets the short-term forecasting needs. The mean value of MAPE is 0.384 and less than 0.5. And further find that the ARIMAX model predicts the effect of multivariate on atmospheric pressure very well. Compared with AIC, the penalty term of BIC is larger. Considering the number of samples, the number of samples is large, which can prevent the model complexity from being too high due to the high accuracy of the model.

It can be seen that the degree of freedom in the model fitting reaches 15, indicating that the sample data provides more information. Because the significance is 0.121 greater than 0.05, there is no statistical difference. On the one hand, compared with other time sequence models, the ARIMAX model performs more stably when dealing with time sequence, and can give better prediction results for different sequences, and the ARIMA model can accurately predict the future trend and periodic changes of time sequence. On the other hand, compared with the ARMA and ARIMA models, which are limited to the prediction of a single sequence, the ARIMAX model adds exogenous variables to analyze the impact of exogenous variables on the original time sequence, making the predicted values make up for some defects of the ARIMA model which is limited to short-term prediction.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Liu Lidan. Weather prediction study based on a probability map model. Nanjing University of Aeronautics and Astronautics, 2019.

[2] Wang Yanpeng, Chen Yan, Yin Huimin et al. Numerical simulation of dust weather field in northern China. Drought Weather, 2007, (04): 12-17.

[3] Zhou Chang, Li Yepei. The UK offshore wind power environmental review mechanism. Proceedings of the 2017 National Environmental and Resources Law Seminar, 2017: 1100-1105.

[4] Liu Haizhen, Zhen Shufang, Wen Xiaohui, et al. Study on the influence of ambient temperature. Modern Electronic Technology, 2023, 46(10): 21-25.

[5] Qiao Nian, et al. CMIP 6 model the seasonal cycle of atmospheric central and southern hemisphere. Journal of Atmospheric Science, 2022, 45(06): 890-903.

[6] Wei Chuorui, et al. Study on the spectral characteristics of pressure pulsation and turbulent atmosphere. Journal of Peking University (Natural Science Edition), 2022, 58(01): 186-194.

[7] Lu Shengdong, et al. Study on the influence of meteorological factors on air quality in Taiyuan area. Desert and Oasis Weather, 2021, 15(02): 98-105.

[8] Chen Jinbei. Application of linear thermodynamics of nonequilibrium states in atmospheric turbulence. Lanzhou University, 2006.

[9] Vincent E M, et al. Influence of upper ocean stratification interannual variability on tropical cyclones. Journal of Advances in Modeling Earth Systems, 2014, 6(3): 680-699.

[10] Li Jiawen. Application of fuzzy treatment of air pressure characteristics in predicting spring cold weather. Guangxi Meteorological, 1992(04): 10-12.

[11] Wu Jian, et al. CMIP 6 simulation evaluation and future estimation of climate change in the Yangtze River Basi. Resources and Environment in the Yangtze River Basin, 2023, 32(01): 137-150.

[12] Liu Chunlin, Wang Tao, Liu Pengyu. Nonlinear correction algorithm for air pressure sensor based on RBF. Meteorological and Environmental Science, 2023, 46(03): 106-111.

[13] Gen Shanshan, et al. Reicability analysis of ERA5 sea surface pressure and wind speed in Bo Sea and North Yellow Sea. Ocean Bulletin, 2023, 42(02): 159-168.

[14] Xu Siyi, et al. An improvement model of global atmospheric air temperature was constructed using the Fourier function. Surveying and mapping science, 2022, 47(11): 170-176.