

Team Control Number

12598

Problem Chosen

B**2022**

HiMCM/MidMCM

Summary Sheet

Prediction of carbon dioxide content and temperature based on function fitting and time series

Summary

Since the Industrial Revolution, the level of carbon dioxide in the atmosphere has kept rising continuously which has resulted in series of climate changes. In this paper we propose a reasonable fitting model of CO₂ concentration level and land-ocean temperature, and show the linear fluctuation relation between them.

Response to question 1: to obtain the overall impact of the change in carbon dioxide in March 2004, the carbon dioxide data from one given year to the previous decade are fitted linearly. The slope obtained from the fitting of 2004 to the previous decade is 1.853, which is larger than the fitting data at any previous time period. The growth value and the average growth rate of 2004 are 2.076 and 0.539% respectively, which are higher than any previous year. Taken together, the change in CO₂ in March 2004 resulted in a larger increase than that has been observed in any previous decade. For the level of CO₂ prediction, we establish the linear regression model, quadratic function model, logarithmic function model, Logistic model, and ARIMA time series to fit the CO₂ content data. The mean square error of them are, respectively, 15.3758, 0.5185, 16.3140, 0.4798 and 0.2743 respectively. After proper analysis, the ARIMA model is the most accurate.

Response to question 2: according to the conclusion of question one, the ARIMA model is chosen to analyze and forecast the future land-ocean temperature in time series, and the Pearson correlation analysis is used to verify the Linear independence between the two models. A Pearson product-moment correlation coefficient of 0.9613 indicates a high degree of Linear independence, and a land-ocean temperature fluctuation model of CO₂ levels reflects the relationship between the two. The wave function modulus is obtained by least squares fitting function:

$$DAP_t = -3.34 + 0.01P_t - 0.053 \cos(0.51P_t + 19.9255) - 0.050 \sin(0.20P_t - 2.54) .$$

To obtain the accuracy of the model prediction, the carbon dioxide concentration level predicted by the ARIMA model is substituted into the fluctuation model, and the temperature prediction results of the fluctuation model are obtained. The results of the 95% confidence interval of the land-ocean temperature predicted by the ARIMA model are regarded as reference values. The deviation rate of the fluctuation model is calculated, and the prediction results of the fluctuation model for the next 45 years are reliable. Since carbon dioxide is not the only greenhouse gas, we consider the linearity of methane to temperature effect, a primary independent variable of methane concentration is added to the re-fluctuation model to obtain the fluctuation model:

$$DAP_t = -4.0412 + 0.0078P_t - 0.0537 \cos (0.5054P_t + 19.9255) - 0.0474 \sin (0.2036P_t - 2.5447) + 0.00093.$$

The advantages and disadvantages of the model are summarized.

Keywords: polynomial fitting Logistic model ARIMA time series model, wave function mode

Table of contents

1.The presentation and restatement of the Question 1	
1.1 Background to question.....	1
1.2 Restatement of Question.....	1
2.Problem Analysis	
2.1 The overall analysis flow chart.....	2
2.2 Analysis of Question 1.....	2
2.3 Analysis of Question 2	
3. Model hypothesis 2	
4. Symbol Illustration 3	
5. The establishment and solution of the Model 3	
5.1 The establishment and solution of Problem-1 Model.....	3
5.1.1 Establishment of Problem-1 Model.....	3
5.1.2 Problem-the solution of the Model.....	5
5.2 Problem two model establishment and solution.....	12
5.2.1 Establishment of Problem-1 Model.....	12
5.2.2 Problem-a model solution	14
6. Non-technical articles 21	
7.Analysis and test of the Model 22	
8.Sensitivity analysis of deviation rate of fluctuation model 22	
9. Evaluation and improvement of the Model 22	
9.1 Advantages of the Model.....	22
9.2 Disadvantages of the Model.....	22
9.3 Improvements to the Model.....	22
9. Reference. 23	
Appendix	

1 The question and restatement

1.1 The research background

Carbon Dioxide and global warming are hot issues in environmental protection. Due to the industrial revolution, atmospheric carbon dioxide concentrations have increased from around 280 ppm to around 421 ppm. Increased atmospheric carbon dioxide contributes to global warming through the greenhouse effect.

1.2 A restatement of the question

The editors of *Science Today* provided the CO₂ concentration level and temperature data sets and asked our team to address current and future projections of CO₂ concentration levels by asking the following questions:

1. Do you agree with the carbon dioxide level statement? Use CO₂ concentration data set 1 to analyze changes in CO₂ concentration.

a. Do you agree that the March 2004 increase in CO₂ led to a larger increase than has been observed in any previous 10-year period? Give relevant explanations.

b. Use more than one mathematical model to fit the data, describing past levels of carbon dioxide and predicting future levels of carbon dioxide in the atmosphere.

c. Use each of the models given above to predict atmospheric CO₂ concentration in 2100. Does any model projection agree that CO₂ concentration level will reach 685 ppm by 2050? If 685 ppm cannot be reached by 2050, when will the model's prediction result reach 685 ppm?

d. Choose the most accurate model and give reasons.

2. Show the relationship between global warming and carbon dioxide based on the data and compare land-ocean temperatures and carbon dioxide concentration levels.

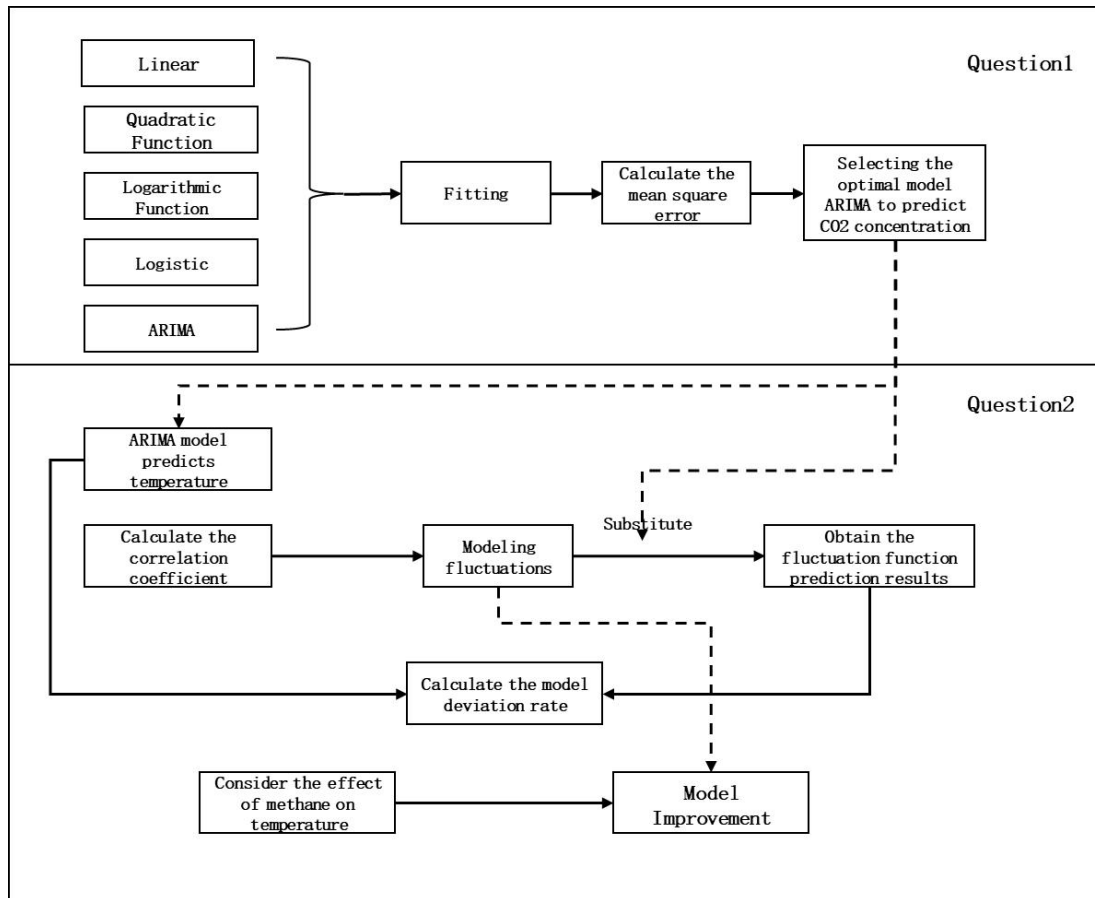
a. Build models to predict future land-ocean changes, and use the models to predict when the average land - ocean temperature changed by 1.25 °C, 1.50 °C and 2 °C? Compared with the base period of 1951-1980.

b. Try to build a model to analyze the relationship between CO₂ concentration levels and land-ocean temperatures since 1959 and explain the relationship or prove that there is no relationship.

c. Extend the model from 2.0. Part B to the future. How far in the future will the model be reliable? If so, what needs to be taken into account for the models to predict future levels of carbon dioxide and land-ocean temperatures?

3. Write a non-technical article for *Science Today* explaining to your team their findings and possible recommendations for the future.

2. Problem analysis



2.1 Analysis of problem 1

For *a*, to relate the 2004 CO₂ concentration levels to the magnitude of the increase observed over any previous 10-year period, by using the data provided in the Annex, 1968~2021 as the year of termination, the slope is obtained by linear fitting in 54 groups, and the effect of the increase of carbon dioxide in March 2004 on the increase of carbon dioxide is obtained. A linear model, a quadratic function model, a Logistic model and a time series model are presented to fit the data and predict the atmospheric carbon dioxide concentration in 2100. For *d*, the data of fitting degree and root mean square error of all models were compared, and the most accurate model was selected by appropriate analysis.

2.2 Analysis of problem 2

For *a*, the ARIMA model is used to predict the future land-ocean temperature changes due to the fluctuating rising trend of temperature, and the final prediction time of the corresponding temperature changes is given for *b*. The correlation between carbon dioxide and land-ocean temperature is analyzed, and the wave function with sine and cosine function is established by taking carbon dioxide as independent variable and land-ocean temperature as dependent variable. For *c*, due to the lack of reliable data, we believe that the ARIMA model is more accurate than the previous model, so we first use the ARIMA (3,2,2) model of CO₂ concentration to predict CO₂ concentration levels 100 years after 2021, and then put the predicted carbon dioxide level into the fluctuation function of temperature with respect to the carbon dioxide concentration level, and the prediction result of the fluctuation function is

obtained. The land-ocean temperature ARIMA (2,1,5) model is used to obtain the prediction of temperature and the final calculation of the deviation rate of the prediction of the wave function. Considering that carbon dioxide is not the only factor affecting temperature, a primary term independent variable of methane concentration can be added to the fluctuation model.

3. Model Assumptions

Assuming1: that the data collected in this paper are true and reliable, which meet the basic requirements of data classification and analysis, accurately reflect the basic laws of global climate change.

Reason1: data analysis rules out data anomalies and so on.

Assuming2: that the energy mix remains the same for decades to come, there will be no major breakthroughs in technology.

Reason2: make sure you can reasonably predict the future based on what you know so far.

4. The explanation of symbols

5.

Symbols	Discription	Units
t	Year	/
P_t	Carbon dioxide concentration level	ppm
k_i	The coefficient obtained by fitting	/
t_0	The initial year	/
μ	ARIMA model constant	/
p	Order	/
γ_i	Autocorrelation	/
ε_t	Error	/
MSE	Mean Square error	/
$\rho_{x,y}$	The correlation between the X and Y variables	/
P_t	T-year CO2 concentration	ppm
D_{P_t}	Land and ocean temperatures at PT for carbon dioxide	°C
DA_{P_t}	Wave models predict land-ocean temperatures	°C
δ_t	T-year fluctuation model forecast deviation rate	/
D_t^{max}	The upper bound of 95% confidence interval of t-year temperature predicted by time series	°C
D_t^{min}	The lower bound of 95% confidence interval of t-year temperature predicted by time series	°C

5. The establishment and solution of the model

5.1 The establishment and solution of Problem 1 model

5.1.1 The establishment of the model

To relate the 2004 CO2 levels to the magnitude of the increase observed over any previous 10-year period, the slope of least squares data from one year to the previous decade is used to determine the year's impact on carbon dioxide levels.

To predict CO2 levels over the coming decades, the data are fitted by using the following models:

Linear model

$$P_t = k_1 t + k_2 \quad (1)$$

P_t is the carbon dioxide concentration of T year, t is the year, k_l is the constant.

$$P_t = k_1 t^2 + k_2 t + k_3 \quad (2)$$

Quadratic function model

P_t is the carbon dioxide concentration of t year, t is the year, k_l is the constant.

Log model

$$P_t = k_1 \ln t + k_2 \quad (3)$$

P_t is the carbon dioxide concentration of t year, t is the year, k_l is the constant.

Improved Logistic model

Considering the human intervention and other factors, the carbon dioxide content may eventually tend to be at a certain value, so the S-type Logistic growth model can be established, and the traditional Logistic curve model has poor fitting effect, without considering the influence of some random factors. The predicted value is often far from the true value, and it is difficult to meet the demand of prediction. The improved model is:

$$P_t = \frac{1}{k_1 + k_2 e^{-k_3(t-t_0)}} + k_4$$

P_t is the carbon dioxide concentration, t is the year, t_0 is the initial year, k_i is the constant.

ARIMA auto-regressive time series prediction model

Time series analysis and research time series in the long-term change in the process of statistical regularity.

ARIMA autoregressive time series prediction model is a combination of difference model, autoregressive model and moving average model:

$$\text{ARIMA}(p,d,q) \quad (5)$$

The difference model is to make the data more stable, the difference order is represented by Auto regressive model

The auto regressive model uses the historical data of factors to forecast itself, and the prediction can only be carried out when the requirement of smoothness is satisfied. The smoothness is checked by Augmented Dickey-Fuller(ADF).

The autoregressive model is expressed as follows:

$$P_t = \mu + \varepsilon_t + \sum_{i=1}^p \gamma_i y_{t-i} \quad (6)$$

y_t is the current value, μ is the constant, p is the order, γ_i is the Auto correlation, and ε_t is the error

Moving Average Model MA

The moving average model is used to eliminate the random fluctuation of the autoregressive model, that is, the error ε_t , which is expressed as follows:

$$P_t = \mu + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (7)$$

Auto-regressive moving average model ARMA

The auto-regressive moving average model is obtained by combining the auto-regressive model and the moving average model:

$$P_t = \mu + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \sum_{i=1}^p \gamma_i y_{t-i} \quad (8)$$

According to the ARIMA (p, D, Q) modeling idea of Box-Jenkins, some necessary verification work is required:

Stationary test of time series data.

The unit root test is needed to judge the smoothness of the time series. In this paper, we use Augmented Dickey -Fuller (ADF) test. The original hypothesis is: if there is unit root, the sequence is not stable. If the original hypothesis is accepted, the sequence is unstable; If the original hypothesis is rejected, the sequence can be considered stable.

White noise test for residual sequence

The main test is whether the residual sequence res is White noise or not and whether there is self-correlation. The assumption is that there is no auto-correlation, that it is White noise. If the original hypothesis is accepted, the RES is considered to be a White noise; If the original hypothesis is rejected, the RES can be considered to be White noise. If the residuals are White noise, the useful information in the time series is totally extracted, and only the random disturbance data are left which cannot be used reasonably. If the residuals are not White noise, then the useful information in the residuals still remains to be extracted, and the model needs to be modified or further extracted.

Mean square error

By comparing the mean square errors of the models, the index calculates the mean of the sum of squares of the errors of the fitting data and the corresponding sample points of the original data, the smaller the value, the better the fitting effect.

$$MSE = \frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{m} \quad (9)$$

MSE is the mean square error, m is the number of samples, y_i is the original data, and $\hat{y}_i$ is the fitting data.

5.1.2 The solution of problem-a model

The slope of the least squares from one given year to the previous 10 years can be used to estimate the effect on CO2 levels of that year. The slope is shown in the following figure:

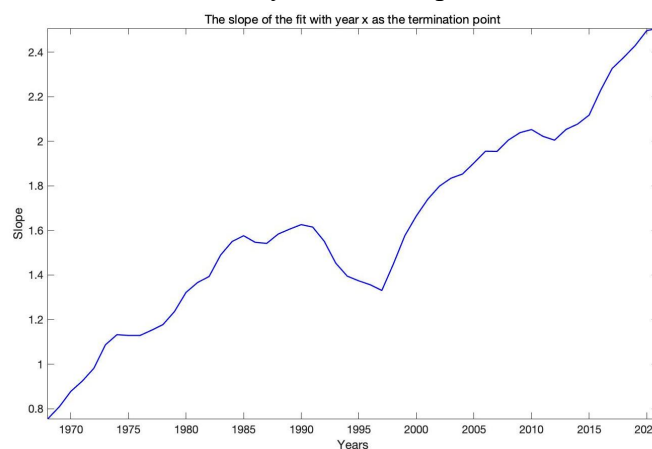


Figure 2: slope from linear fitting of ten-year data

Data from 2004 for the decade before the end of the year are fitted with a steeper slope than

data from any previous year

The fitted slopes are larger, suggesting that the March 2004 increase in CO₂ led to a larger increase than that has been observed over any previous 10-year period.

The average growth rate and the average growth rate in the following ten years can reflect the influence of the change of the carbon dioxide concentration level in that year on the following period.

In fact, there is little difference between the average value of growth and the average value of growth rate in the last ten years from 2001 to 2004, but there is still an upward trend. Changes in CO₂ concentration levels have resulted in larger increases than those have been observed over any previous 10-year period.

(1) Linear function fitting

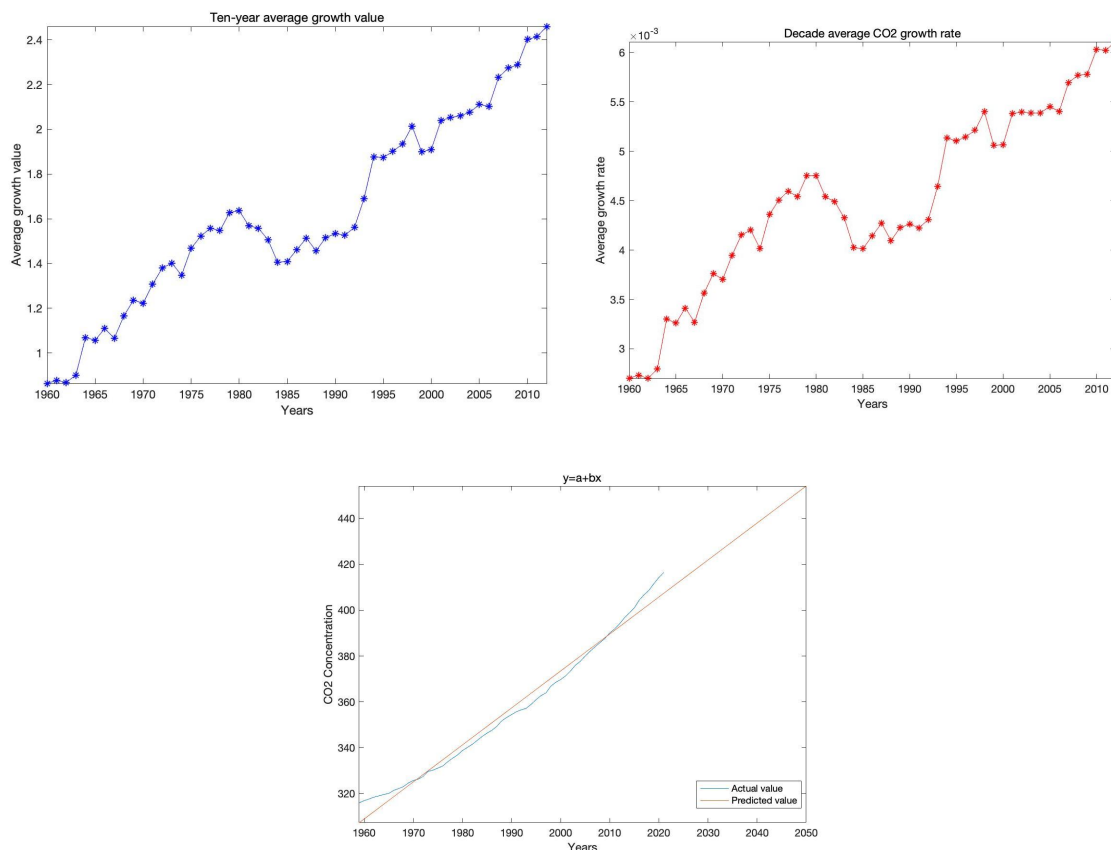
The relationship between carbon dioxide concentration level and time obtained by linear fitting is as follows:

$$P_t = 1.6140t - 2854.59 \quad (10)$$

P_t is carbon dioxide concentration level, t is year.

The linear goodness of fit is 98.25% . It is predicted that the CO₂ concentration will reach 454.1807 ppm in 2050, 534.8825 ppm in 2100, and 685 ppm in 2193 ~ 2194. The mean square error is 15.3758.

The average of the growth and the average of the growth rate over a decade starting from a given year are shown in the following chart:



(2) Quadratic function fitting

The relationship between time and CO₂ concentration level obtained by quadratic function fitting is as follows:

$$P_t = 0.0130t^2 - 50.2758t + 48771.4 \quad (11)$$

The goodness of fit of quadratic function is 99.94%. It is predicted that the level of carbon dioxide concentration will reach 496.8051 ppm in 2050 and 688.3268 ppm in 2100, the quadratic function predicts that the CO₂ concentration will reach 685 ppm in 2099-2100, with a mean square error of 0.5185.

(3) Lognormal function fitting

The log-function fit shows the following relationship between CO₂ concentration and time:

$$P_t = 3210\ln t - 24025.29 \quad (12)$$

Lognormal function fitting predicts that the CO₂ concentration will reach 452.826 ppm in 2050, 530.179 ppm in 2100, and 685 ppm in 2203-2204, the mean square error is 16.3140.

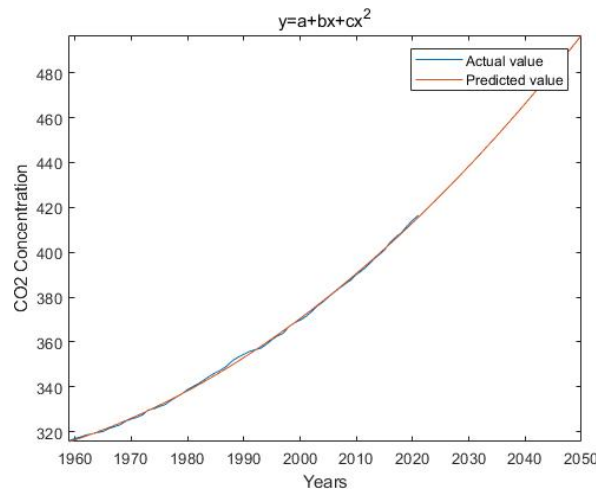


Figure 6: quadratic function fitting results

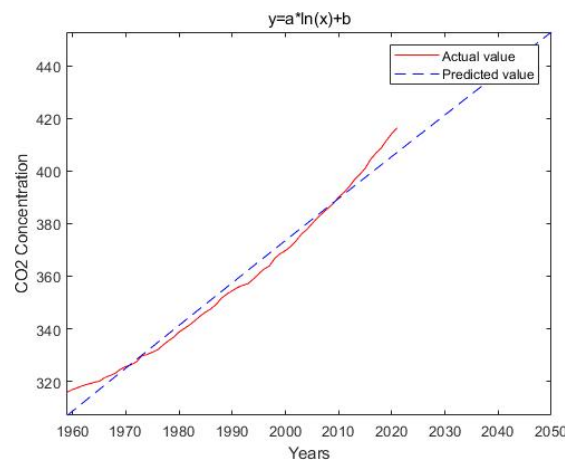


Figure 7: log function fit results

It is seen from the figure that the result of log function fitting is similar to linear effect in a short time.

(4) The improved Logistic model

The relationship between CO₂ concentration level and time obtained by Logistic model fitting is as follows:

$$P_t = \frac{1}{0.000396 + 0.0179e^{0.0176(t-t_0)}} + 260.0181 \quad (13)$$

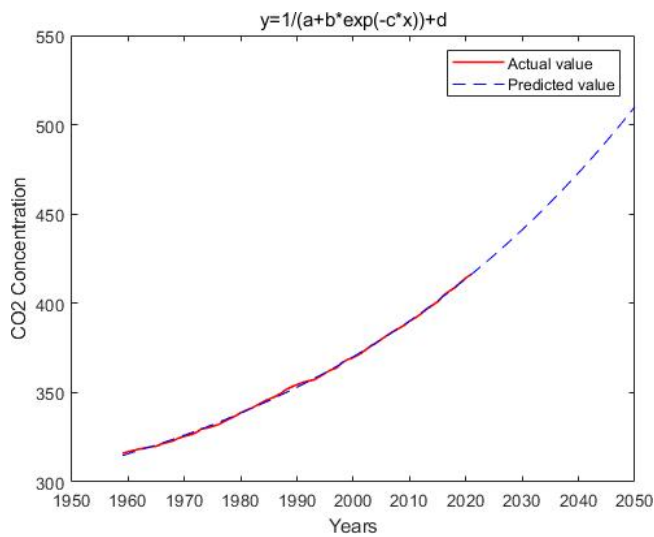


Figure 8: fitting results of S-type growth model

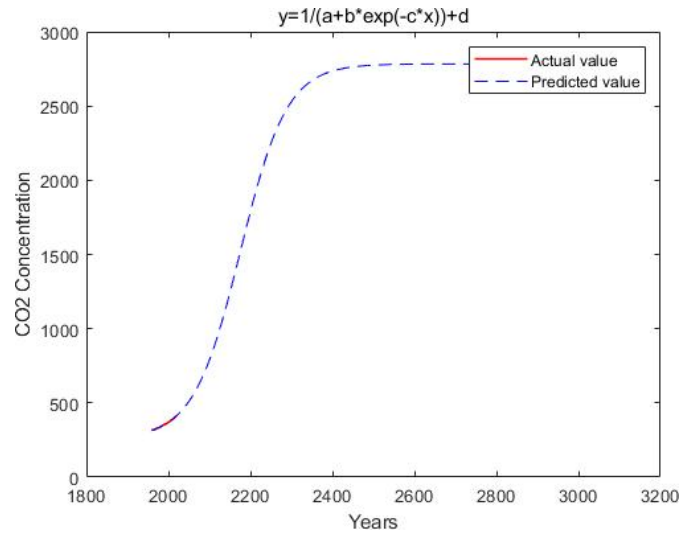


Figure 9: s-shaped growth

The logistic growth model predicts that the CO₂ concentration will reach 486.6149 ppm in 2050, which means CO₂ concentration does not reach 685ppm in 2050. It will reach 703.1663 ppm in 2100 and 685 ppm in 2097-2098. The mean square error is 3.3610. The s-shaped growth function as a whole is shown in the figure9.

(5) ARIMA autoregressive time series prediction model

The steps are as follows by using Box modeling:

Step 1. The unit root test for the smoothness of the original sequence: if the sequence is not smooth, make the sequence smooth through the difference transformation or other transformation.

Step 2. The order p and q of ARMA model are determined by correlation coefficient method or criterion function method.

Step 3: Carry out diagnostic analysis, residual test.

The flow chart is as follows:

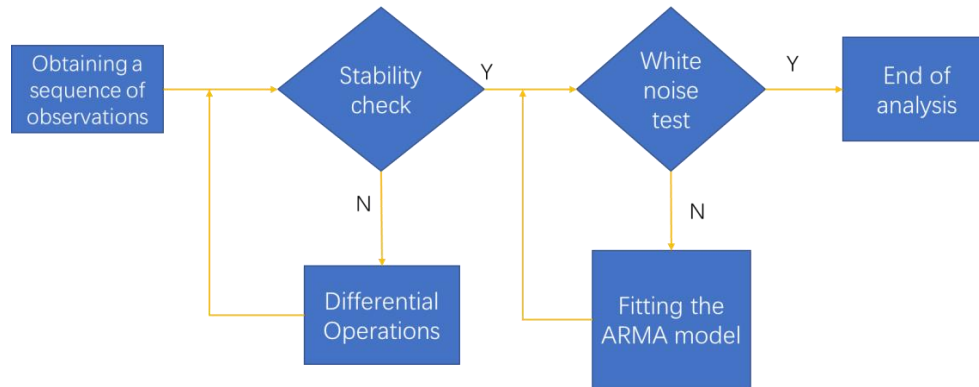


Figure 10: Box Jenkins Flowchart

The fluctuation trend of CO₂ concentration from 1959 to 2021 is plotted, and the stability of the original series is checked. First, the time-varying trends of CO₂ concentrations from 1959 to 2010 are shown in the following chart:

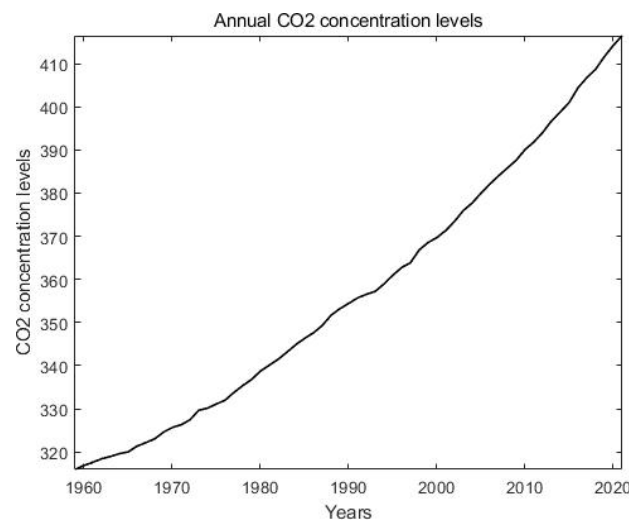


Figure 11: time-varying trends of CO₂ concentrations from 1959 to 2000

The results of the smoothness test on the original data are as follows:

Table 1: Stability Test results

h-Value	P-Value
h-Value	P-Value
0	0.9990

In an autoregressive model such as ARIMA, the stability of the time series data is required. From the results of the original data, the P value of the original series is greater than 0.05, which fails the significance test. Therefore, the original data are not stable, the original data need differential processing.

The first order difference and the second order difference of the original sequence are tested. The results of the first order difference and the second order difference are as follows:

Table 2: Stationarity test for difference data

diff	h-Value	P-Value
First order differential	0	0.2153
Second order differential	1	0.0010

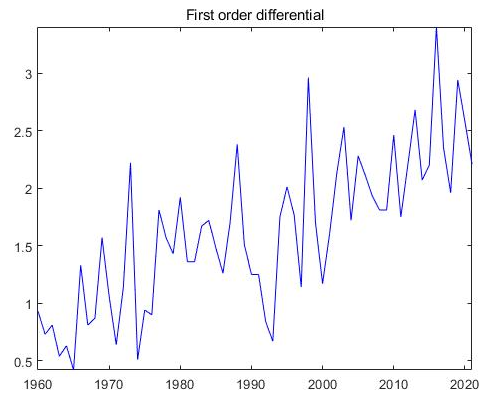


Figure 12: data after First Order Difference

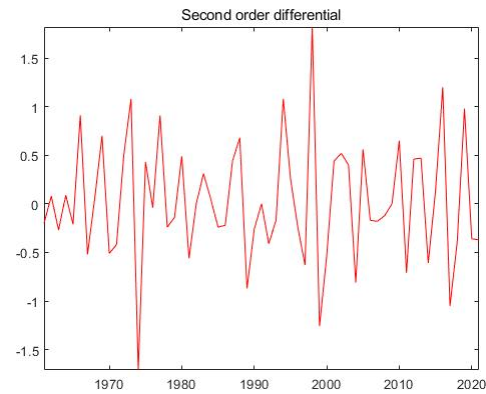


Figure 13: data after second order difference

After the second order difference of the original sequence, the data pass the stationarity test, and then the residual res white noise test is performed. The P value is $0.3367 > 0.05$, and the original hypothesis is accepted that the residual res is White noise, there is no autocorrelation.

From the test results of stationarity and White noise, we can find that the data after the second order difference has passed the test of stationarity and white noise, indicating that it performs to be stable and is well predictable.

Using AIC Akaike information criterion for parameter selection of ARIMA models:

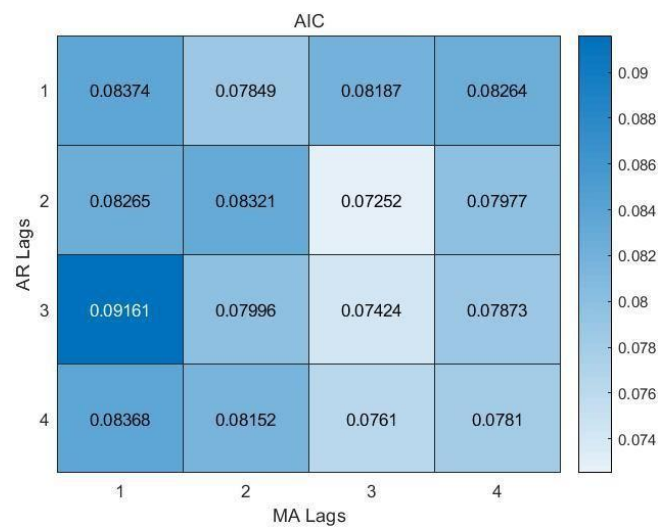


Figure 14: AIC ordering

We can determine $p = 2$, $q = 3$ by AIC criterion, that is, the time series model is ARIMA (2, 2,3) . To achieve the projections for the next 100 years, the results are shown in the following figure:

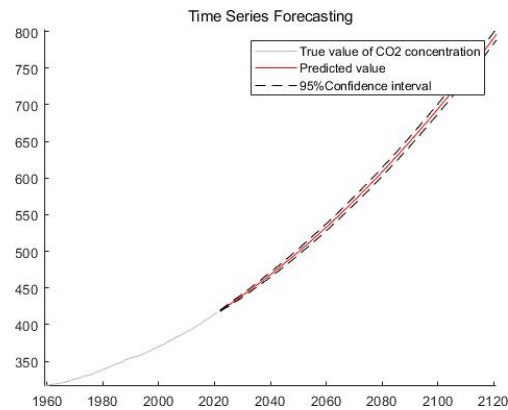


Figure 15: time series prediction results

ARIMA (2, 2, 3) model predicts that the CO₂ concentration in 2100 will be 698.7570. In order to quantify the reliability of the time series, the first 70% of the original data are selected as the training set and the last 30% as the test set, the mean square error is 0.2743.

Comparison of models

Prediction of CO₂ concentration levels in 2100:

Table 3: CO₂ concentration levels for 2100 as predicted by different models

model	2100 CO ₂ concentration(ppm)
Linear	534.8825
Quadratic Function	688.3268
Logarithmic Function	530.1790
Logistic	703.1663
ARIMA	698.7570

Prediction of when CO₂ levels will reach 685 ppm: All of the models predict the time when CO₂ levels achieve 685 ppm to be after 2050. Mean Square error of model predictions:

Table 4: different models to predict when CO₂ levels will reach 685 ppm

model	Achieve 685ppm
Linear	2193~2194
Quadratic Function	2099~2100
Logarithmic Function	2203-2204
Logistic	2097~2098
ARIMA	2097-2098

Table 5: mean square errors of different models

model	MSE
Linear	15.3758
Quadratic Function	0.5185
Logarithmic Function	16.3140
Logistic	0.4798

In data statistics, the mean square error (MSE) is the expected value of the difference between the estimated value and the true value of the parameter.

Although the fitting effect of the linear model is good, the mean square error is big, and the trend of change in different periods has no difference. So it can not realize the effective trend prediction.

The change trend of log-function model becomes smaller in the later period, which is not in accordance with the reality. Moreover the mean square error is big, the fitting effect is bad.

The figure of Logistic function is an s-shaped curve, which can reflect the characteristics of population surge, slow increase and saturation. But when the carbon dioxide concentration tends to saturate, it will be after the year 2300. Therefore the S-type curve has no great fitting significance.

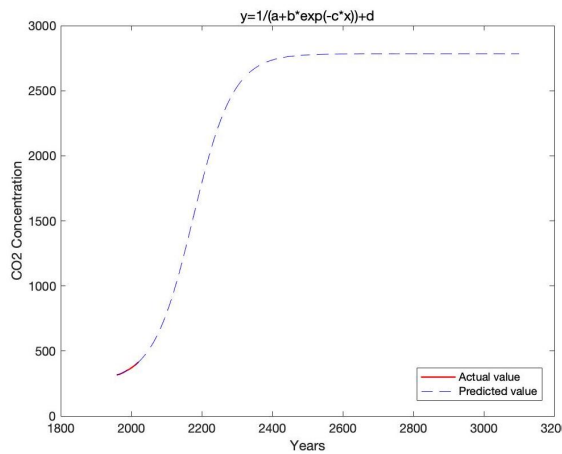


Figure 16: Logistic

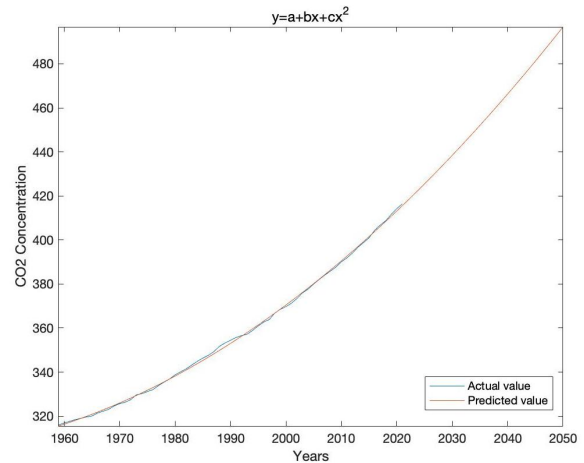


Figure 17: time series prediction

To sum up, we suggest that the effects of linear and logarithmic function fitting are poor. Considering the mean square error, quadratic function is poorer than the Logistic growth model and ARIMA effect, the change trend of quadratic function is absolute and ARIMA model is more accurate than Logistic model. So ARIMA model is the most accurate model in short term.

5.2 Establishment and solution of problem 2 model

5.2.1 Establishment of problem model 2

● Land-ocean temperature prediction model ARIMA

It is necessary to establish a reasonable model to predict the future land-ocean temperature changes. According to the results obtained in question one, ARIMA autoregressive time series model is chosen to predict the temperature.

$$\text{ARIMA}(p, D, Q) \quad (14)$$

1) Auto regressive model AR

$$D_t = \mu + \varepsilon_t + \sum_{i=1}^p \gamma_i Y_{t-i} \quad (15)$$

2) Moving Average Model MA

$$D_t = \mu + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (16)$$

3) Auto-regressive moving average model ARMA

$$D_t = \mu + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \sum_{i=1}^p \gamma_i y_{t-i} \quad (17)$$

According to the ARIMA (p, D, Q) modeling idea of Box-Jenkins, some necessary verification work is required:

Stationary test of time series data

The unit root test is used to judge the smoothness of the time series. Augmented Dickey-Fuller (ADF) test is used in this paper.

The original hypothesis is: if there is unit root, the sequence is not stable. If the original hypothesis is accepted, the sequence is unstable;

If the original hypothesis is rejected, the sequence can be considered stable.

White noise test for residual sequence.

The main test is whether the residual sequence res is White noise or not and whether there is self-correlation. The assumption is that there is no autocorrelation, that it is White noise.

If the original hypothesis is accepted, the RES is considered to be a White noise;

If the original hypothesis is rejected, the RES can be considered to be White noise.

If the residuals are White noise, the useful information in the time series is totally extracted, and only the random disturbance data is left, which cannot be used reasonably.

If the residuals are not White noise, then the useful information in the residuals still can be extracted, and the model needs to be modified or further extracted.

Fluctuation function model for CO2 and land-ocean temperature

First, the correlation between carbon dioxide concentrations and land-ocean temperatures is tested. The Pirsson correlation is calculated by using the following formula:

$$\rho_{x,y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} \quad (18)$$

The Pearson product-moment correlation coefficient is to measure the linear correlation between two variables. The correlation coefficient is close to 1 or -1. The stronger the correlation or negative correlation is, the closer the correlation coefficient is to 0, the weaker the correlation is. Given the strong Linear independence between the two, and given the possible cyclical volatility of the data,

Build a wave function model:

$$DU_{P_t} = k_1 + k_2 P_t + k_3 \cos(k_4 P_t + k_5) + k_6 \sin(k_7 P_t + k_8) \quad (19)$$

Pt is the carbon dioxide concentration of t year, DPt is the carbon dioxide temperature of Pt Land and sea, and ki is the constant.

Expansion of the relationship between carbon dioxide concentration and land-ocean temperature

In order to use the model of CO2 concentration and land-ocean temperature to predict the future, we suggest that the ARIMA model is more accurate, and take the ARIMA model as the reference value.

$$D_t = \mu + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \sum_{i=1}^p \gamma_i y_{t-i} \quad (20)$$

The ARIMA model takes 95% confidence interval to get the temperature interval at each time point:

$$Dt \in [Dtmin, Dtmax] \quad (21)$$

Firstly, the ARIMA model of carbon dioxide concentration is combined with the model of the relationship between carbon dioxide concentration and land-ocean temperature. The results of the ARIMA model for carbon dioxide concentration are more accurate. The ARIMA model for carbon dioxide concentration is as follows:

$$P_t = \mu + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \sum_{i=1}^p \gamma_i y_{t-i} \quad (22)$$

(1) Substitution of the CO2 ARIMA model into the CO2 concentration and land-ocean temperature fluctuation function model:

$$DU_{P_t} = k_1 + k_2 P_t + k_3 \cos(k_4 P_t + k_5) + k_6 \sin(k_7 P_t + k_8)$$

(2) calculation of the deviation rate of the fluctuation function model:

$$\delta_t = \begin{cases} (D_{P_t} - D_t^{max})/D_t^{max}, & D_{P_t} > D_t^{max} \\ 0, & D_t^{min} \leq D_{P_t} \leq D_t^{max} \\ (D_t^{min} - D_{P_t})/D_t^{min}, & D_{P_t} < D_t^{min} \end{cases} \quad (23)$$

(3) If the deviation rate δ_t is less than 25% , the wave function model is considered to be more accurate.

5.2.2 The solution of problem 2 model

● Land-ocean temperature prediction model

Using Box modeling to solve the idea, the steps are as follows:

Step1. Do the unit root test for the smoothness of the original sequence. If the sequence is not smooth, make the sequence smooth through the difference transformation or other transformation.

Step2. The order P and Q of ARMA model are determined by correlation coefficient method or criterion function method

Step3. Perform diagnostic analysis and residual test.

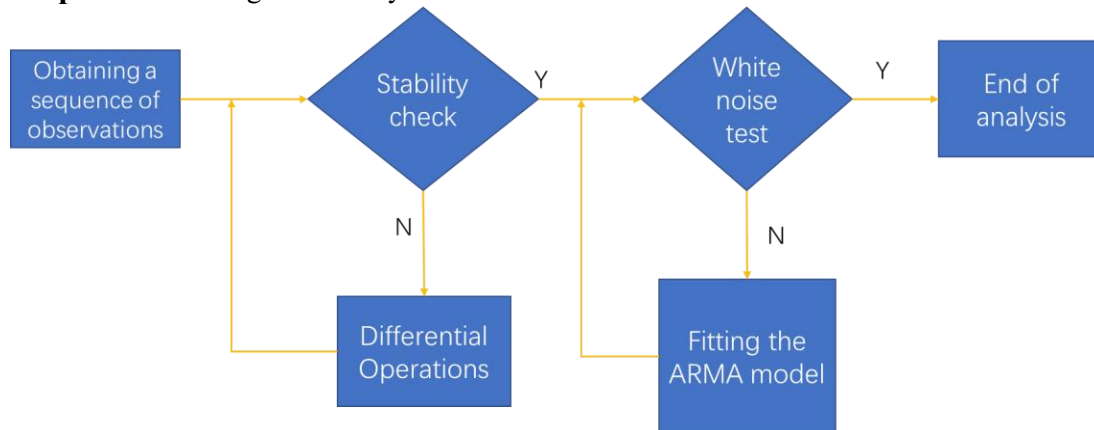


Figure 18: Box Jenkins Flowchart

More detailed data are obtained from the literature, and the fluctuation trend of land-ocean temperature from 1880 to 1880 is plotted, and the 2021 stability of the original series is tested.

First, the time-varying trends of land-ocean temperatures from 1880 to 2021 are shown in the following chart:

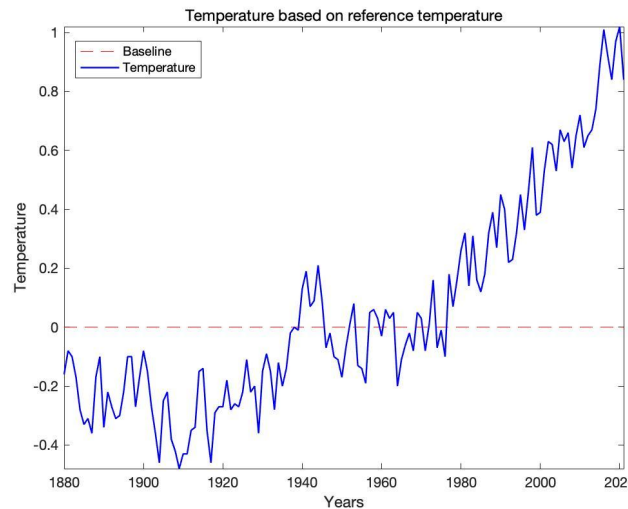


Figure 19: temperature trends

The results of the smoothness test on the original data are as follows:

Table 6: Stationarity Test

h-Value	P-Value
0	0.2366

In an autoregressive model such as ARIMA, the stability of time series data is required.

The result of the original data shows that the P value of the original sequence is greater than 0.05, which cannot pass the significance test.

The first order difference of the original sequence is tested and the results are as follows:

Table 7: stationarity test for difference data

diff	h-Value	P-Value
First order differential	1	0.0010

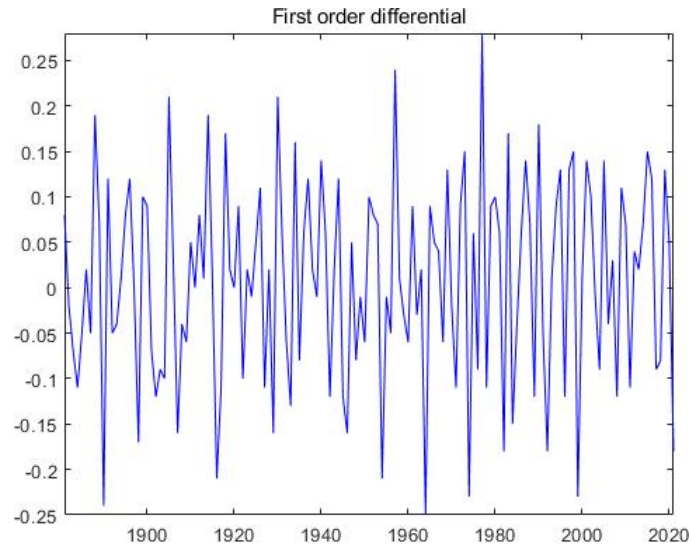


Figure 20: First Order difference

The data after the first order difference of the sequence pass the stationarity test, and then the residual res white noise test,

The p value was $0.3367 > 0.05$. The original hypothesis was accepted and the residual res is considered to be White noise. From the test results of stationarity and white noise, it can be seen that the data after first order difference pass the stationarity and White noise

The test shows that the data after the first order difference is stable and has good predictability. Use AIC Akaike information criterion for parameter selection of ARIMA models:

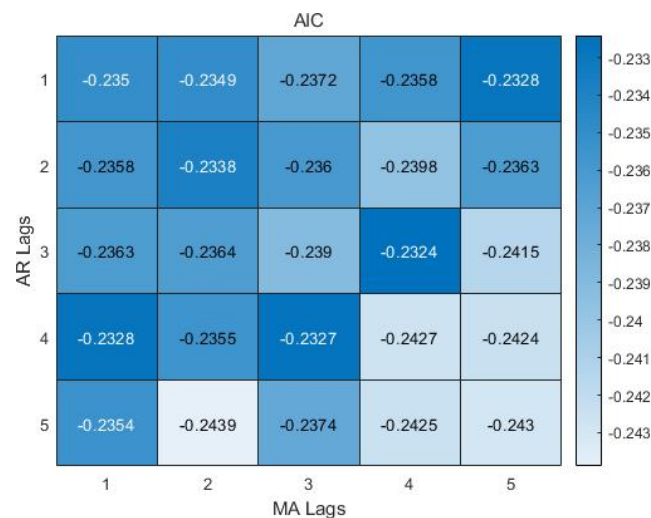


Figure 21: AIC order determination

From the figure we can determine $p = 2$, $q = 5$, that is, the time series model is ARIMA (2,1,5). To achieve the projections for the next 100 years, the results are shown in the following figure:

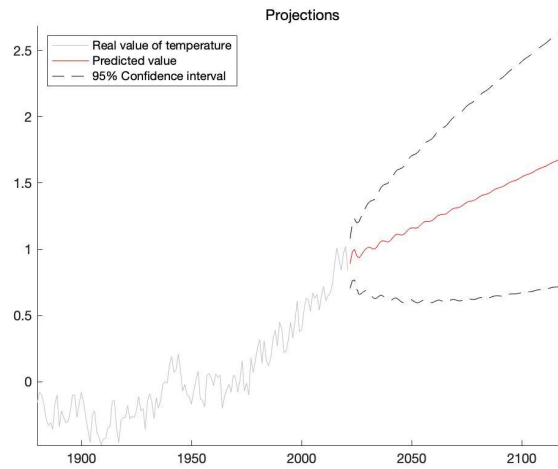


Figure 22: ARIMA predicted temperature

According to the predicted data, the average land-ocean temperature will change by 1.25°C, 1.5°C, 2°C, respectively, in 2070~2071, 2107~2108, 2181~2182.

● Land-ocean temperature fluctuation model of CO2 levels

First, the data of carbon dioxide concentration and land-ocean temperature are normalized separately, and the formula is as follows:

$$X_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (24)$$

The normalized data are as shown in the figure:

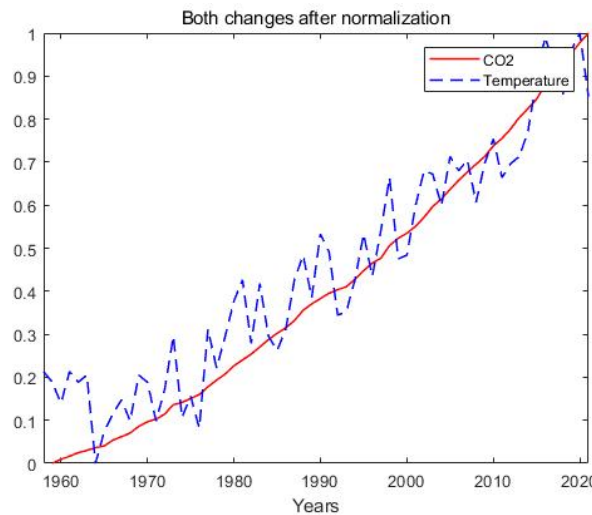


Figure 23: normalized comparison

The figure supposes that there is some linear relationship between the two. To test the linear independence of the two, the Pearson product-moment correlation coefficient is calculated as follows:

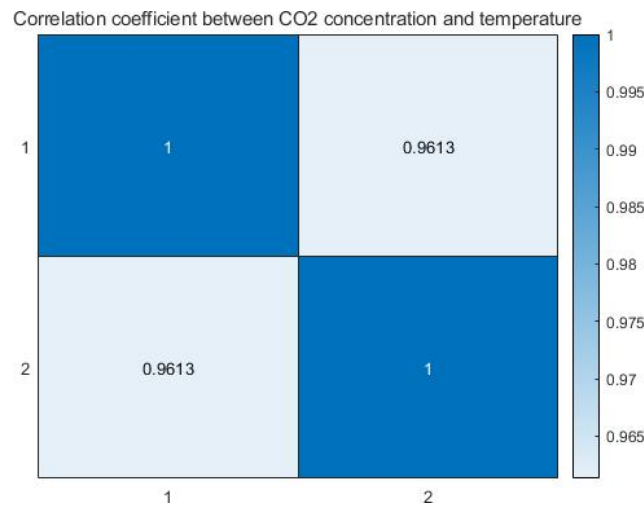


Figure 24: correlation coefficient

There is a strong linear independence between the two, so a wave function can be modeled:

$$DA_{P_t} = k_1 + k_2 P_t + k_3 \cos(k_4 P_t + k_5) + k_6 \sin(k_7 P_t + k_8) \quad (25)$$

by using a least squares fit to produce a wave function:

$$DA_{P_t} = -3.34 + 0.01P_t - 0.0525 \cos(0.5054P_t + 19.9255) - 0.0504 \sin(0.2036P_t - 2.5447) \quad (26)$$

The results of the wave function prediction are shown in the following figure:

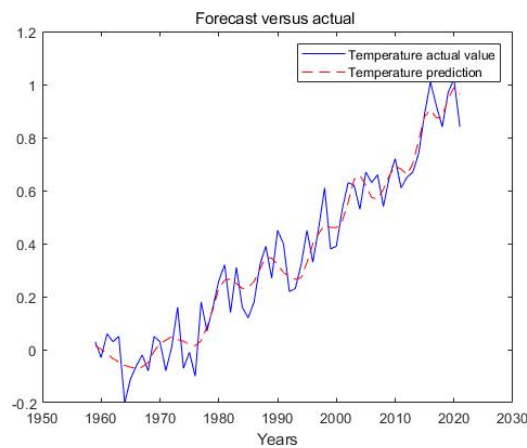


Figure 25: wave function prediction

In order to obtain the reliability of the wave model, we use the results of the time series as a benchmark, the specific steps are as follows:

Step1. The ARIMA (2,2,3) model and the ARIMA (5,1,2) model are used to predict CO2 levels and land-ocean temperatures, respectively, for the next 100 years of 2021.

Step2. The predicted land-ocean temperature of ARIMA model is obtained by substituting the predicted results of carbon dioxide concentration level into the fluctuation function model.

Step3. The deviation rate of the wave function is calculated by comparing the wave function

with the results of time series. The comparison between the time series and wave function predictions is shown in the figure:

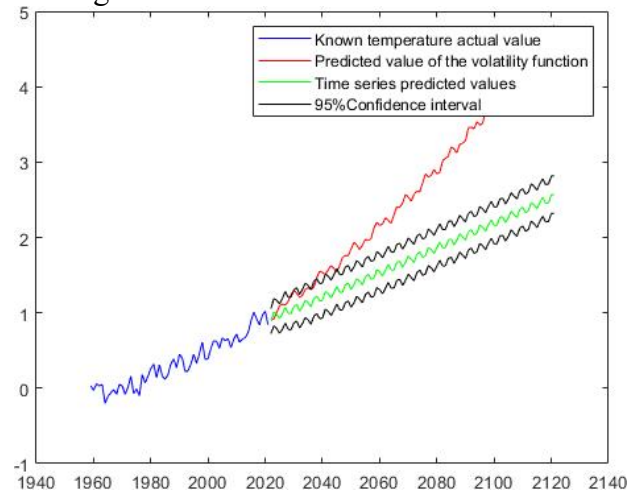


Figure 26: Comparison of predicted results

The minimum deviation rate δt of 0.25 is considered to be reliable for the wave model 2021 to predict the deviation rate 100-year afterwards:

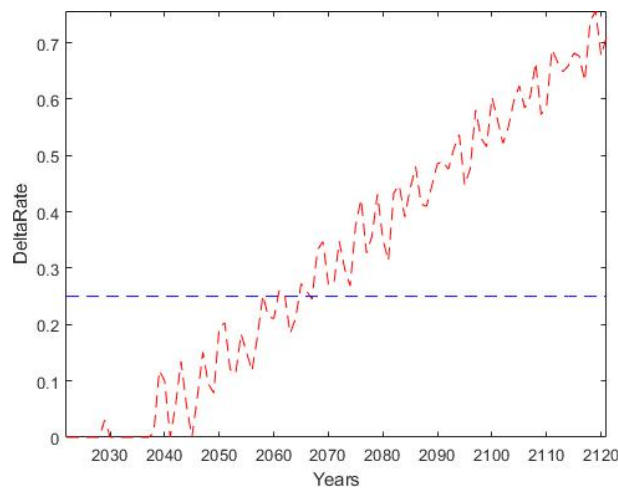


Figure 27: deviation rate

The 2058 predicted deviation rate exceeds 0.25 for the first time and the 2067 predicted deviation rate is considered below 0.25 for the last time

The prediction of wave function model before 2058 is more accurate, and the predicted data from 2058 to 2067 can be used for reference.

The post-2067 figures are less accurate. We consider the deviation rate last below 0.25 for years prior to the data is reliable, that is, the predicted data for the next 45 years are reliable.

● The effect of fossil CO₂ emission on CO₂ concentration

With the concern that concentration of CO₂ is not attributed to time, which does not have practical influence on concentration, we should build a new model concerning the direct attributes of concentration of CO₂ though ARIMA is able to predict concentration of CO₂ with good accuracy. By doing so, we are finding and using the root cause of the change in concentration of CO₂, which equips us with the ability to relate the concentration of CO₂ to practical variables and investigate the degree of importance of the impact that a variable has on the concentration of CO₂. In this way, efficient methods and plans could be made in controlling concentration of CO₂.

In this way, we consider the fossil CO₂ emissions in tons each year. The CO₂ emitted and inhaled by natural species are generally in balance, so it does not affect the concentration of CO₂. However, the fossil are from the materials below the ground, and are burned to emit CO₂ to break the natural balance above the ground. Fossil burning is considerable harm that is exerted by human beings on the jeopardization of CO₂ concentration.

Since the amount of CO₂ that fossil burning emits is so great, we use $\ln(E)$ to diminish the value for the sake of less demanding computational requirements. We use the fluctuation model fitting and get the following model, relating the each-year growth concentration of CO₂ to the fossil emissions directly, and the model shows great linear relationship and also reasonable fluctuation.

$$P_{t2} - P_{t1} = -1.4804 + 4.5295\ln(E) + 2.7566\cos(1.6108\ln(E) + 0.6964) - 0.5884\sin(0.0684\ln(E) + 2.1659) \quad (27)$$

The effect of methane levels on temperature.

Greenhouse gases contain more than carbon dioxide, so the its level is not the only factor that matter. Consider the effect of methane levels on temperature. Adding a first-order term to the fluctuation model for methane concentration levels:

$$DAPtCt = k1 + k2Pt + k3 \cos(0.5054Pt + 19.9255) + k4 \sin(0.2036Pt - 2.5447) + k5Ct \quad (28)$$

Among them CT is methane concentration.

A binary wave function of temperature with respect to carbon dioxide and methane concentrations is obtained by fitting:

$$DAPtCt = -4.0412 + 0.0078Pt - 0.0537 \cos(0.5054Pt + 19.9255) - 0.0474 \sin(0.2036Pt - 2.5447) + 0.00093Ct \quad (29)$$

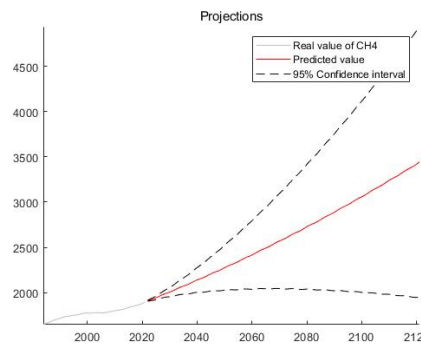


Figure 32: the ARIMA model predicts methane concentrations

The predicted results of carbon dioxide and methane concentrations from the Arima model are put into the improved wave function model, and the predicted results are as follows:

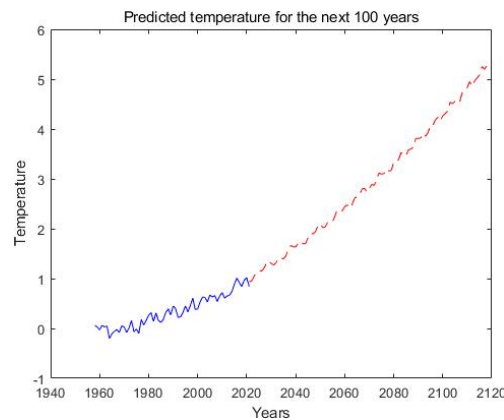


Figure 33: final projections

6. Non-technical article

We use Arima to predict concentration of CO₂ and temperature, and we discover temperature to be fluctuational linear relative to concentration of CO₂ and CH₄. In conclusion, we advocate the factory and people from the globe to use less burning fossil in either daily use or large-scale manufacture. The spread of new and additional new, clean energy use is emergent for the sake of less emission of greenhouse gases such as CO₂ and CH₄.

7. Analysis and test of the model

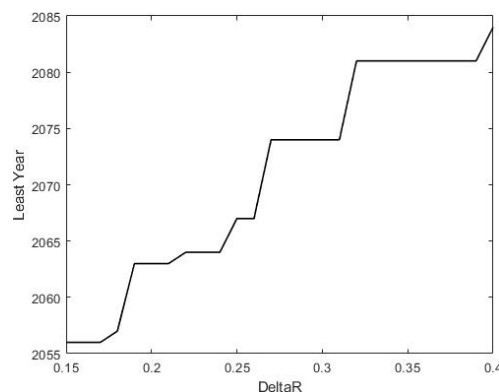


Figure 34: latest reliable years with different minimum deviation rates

The greater the deviation rate, the greater the probability of allowing errors, and the longer the time to determine the reliability of the data prediction.

In fact, because the time series prediction data is only as a reference, the model deviation rate within 0.20 can be connected

It can be considered that the wave function forecast is reliable until 2063.

In the third question, the deviation rate is set artificially and the deviation rate can be changed to know before which year the model prediction data under different deviation rates are reliable. The results are as follows:

8. Evaluation, improvement and promotion of the model

The advantages of the model

1. By using time series analysis and prediction, we can find out the characteristics, trend and developing law of variable change from time series, so as to effectively predict the future changes of variables.
2. The relationship between land-ocean temperature and carbon dioxide concentration is given by establishing a wave function prediction model.

The disadvantages of the model

1. The time series prediction method has the defect of prediction error because the outstanding time series does not consider the influence of external factors.
2. The reference data of the accuracy of the fluctuation relation model is the result of the time series prediction, and it has a certain subjectivity to the value of the deviation rate.

9. References

1. Yu Yong-tao, Zhang Yang. Distribution network load forecasting based on improved logistic model method [J] . Journal of Nanchang University (engineering edition) , 2022,44(02) : 200-204. DOI: 10.13764/J. CNKI. NCDG. 2022.02.006.
2. Shen Chaoqun. Study on saturated load forecasting based on load density method and Logistic model method [D] . Zhengzhou University, 2016.
3. Lin Xujie, Meng Xiangrui. Study on gas concentration prediction based on ARIMA time series [J] . Journal of the Heilongjiang Institute of Technology (integrated edition) , 2022,22(07) : 77-83. DOI: 10.16792/J. CNKI. 1672-6758.2022.07.011.
4. Jarubin, Gao Jinfeng. Time series prediction method of dissolved gas content in transformer oil based on ARIMA Model [J] . Acta Zhengzhou University, 2020,41(02) : 67-72. DOI: 10.13705/J. ISSN. 1671-6833.2020.03.010.

Appendix

Appendix 1
<pre> clc;clear;close all; %import data C02=xlsread('2022_HiMCM_Data.xlsx','C02 Data Set 1'); years=C02(:,1);DataC02=C02(:,2); %dipect graph: C02 concentration level figure(1); plot(years,DataC02,'k-','linewidth',1.2);axis tight; title('Annual C02 concentration levels');xlabel('Years');ylabel('C02 concentration levels'); %% a %fit each decade's data(some year as termination) K=[years(10:end)]; R=K; for i=1:length(years)-9 x=[ones(10,1),years(i:i+9)]; y=DataC02(i:i+9); [b,bint,r,rint,stats]=regress(y,x); K(i,2)=b(2); R(i,2)=stats(1); </pre>

```

end
%diplot graph: slope
figure(2);
plot(K(:,1),K(:,2),'b-','linewidth',1.2);axis tight;
title('The slope of the fit with year x as the termination
point');xlabel('Years');ylabel('Slope');

DiffData=[years(2:end),diff(DataC02)];           %difference
DiffData(:,3)=DiffData(:,2)./DataC02(1:end-1);%growth rate
for i=1:length(DataC02)-10
    DiffData(i,4)=mean(DiffData(i:i+9,2));
    DiffData(i,5)=mean(DiffData(i:i+9,3));
end
%average growth
figure(3);
plot(DiffData(1:end-9,1),DiffData(1:end-9,4),'b*-');axis tight;
title('Ten-year average growth value');xlabel('Years');ylabel('Average
growth value');
%average growth rate
figure(4);
plot(DiffData(1:end-9,1),DiffData(1:end-9,5),'r*-');axis tight;
title('Decade average CO2 growth rate');xlabel('Years');ylabel('Average
growth rate');
%% linear fitting y=a+bx
x=[ones(size(years)),years];y=DataC02;
[b,bint,r,rint,stats]=regress(y,x);
X=C02(1):2050;Y=b(1)+b(2)*X;

figure(5);
plot(years,y,X,Y);axis tight;
title('y=a+bx');xlabel('Years');ylabel('CO2 Concentration');
legend('Actual value','Predicted value');

%mean square error
ytest=DataC02;ytest_fit=b(1)+b(2)*years;
test_error=ytest-ytest_fit;
test_mse1=mse(test_error);

%when to achieve 685 ppm
syms x;
x685=double(solve(b(1)+b(2)*x==685,x));
disp(['linear goodness of fit:',num2str(stats(1))]);
disp(['linear predicted concentration of CO2 in
2050:',num2str(b(1)+b(2)*2050),...
',concentration predicted in 2100:',num2str(b(1)+b(2)*2100)]);
disp(['linear predicted concentration of CO2 will be,
in',num2str(x685),'year, achieve 685 ppm']);

%% quadratic function fitting
x=[ones(size(years)),years,years.^2];y=DataC02;
[b,bint,r,rint,stats]=regress(y,x);
X=C02(1):2050;Y=b(1)+b(2)*X+b(3)*X.^2;

```