Team Number:	2212335
Problem Chosen:	C

2022 APMCM summary sheet

Global climate warming is a phenomenon related to nature. In this paper, the ARIMA model and LSTM model are used to predict the future global average temperature change. The prediction results show that the future global temperature will show a slow rising trend. Further, various factors that may affect the rise in temperature are analyzed by using the Grey Relation Analysis（GRA）, and it is determined that the area of agricultural land has the greatest impact on the rise in temperature.

For question 1, use the data provided by the question to restore the original standard temperature data. In the discussion of (a), it is believed that the increase in temperature in March 2022 is significantly greater than that in the past ten years, because the temperature increase in the previous ten years was 1.1%, and the increase from March 2021 to March 2022 is 1.92%.

In (b), ARIMA model and LSTM model are established to predict the future air temperature respectively. Both models predict a slow rise in temperature.

In the discussion of (c), it is calculated that the ARIMA prediction result is 9.6147 degrees in 2050 and 9.9985 degrees in 2100; the LSTM prediction result is 8.9581 degrees in 2050 and 9.2975 degrees in 2100.

In (d), it is considered that the ARIMA model is better than LSTM, because the p-value of ARIMA is lower, and the RMSE value of LSTM is higher than ARIMA, indicating that the goodness of fit of ARIMA is higher, so we choose the ARIMA model.

For question 2, in (a), using the data set provided by the title and other data sets, it is found that the global average temperature has a significant correlation with the location, and the global land average temperature is higher than the global average temperature, the global land average temperature is 0.6 degrees higher than the global average temperature; the global average is less relevant to location.

In (b), analyze the factors of volcanic eruptions, forest fires and the COVID-19. We find that there is a positive relationship between global temperature and some disasters.

In (c), relevant factors are collected including: forest area, CO_2 emissions, macroeconomic indicators, population, total production, total trade, agricultural land area.

In (d), we need to find some measures to curb or slow down global warming. We think there are 3 measures to curb global warming. The first is to control population growth. Effective control of population growth reduces CO_2 emissions. The second is to plant trees, so that forests can play a greater role in absorbing and transforming CO_2. The third is to improve the utilization technology and utilization rate of various energy sources, make more efficient use, and reduce CO_2 emissions.

For question 3, we write a non-technical article to explain that global warming is getting worse since 2020. With the increasing of activities of human being, global temperature turns into a new high level. We need to slow down activities and lessen the release level of CO_2. Improve utilization technology is a very good way to relieve global warming.

Keywords: Autoregressive integrated moving average　Long short term memory algorithm
Grey relative analysis

Contents

I. Introduction

In order to indicate the origin of problems, the following background is worth mentioning.

Since this year, we have seen a large number of amazing temperature reports. There is no doubt that the earth is burning. Following the terrible high temperature in these areas from the end of June to the beginning of July, Italy set a European temperature record again, reaching an astonishing 48.8 ° C, and many countries declared a state of emergency.

Global warming is a phenomenon related to nature. It is precisely because of the continuous accumulation of the greenhouse effect that the energy absorbed and emitted by the Earth's atmospheric system is unbalanced. The continuous accumulation of the energy of the Earth's atmospheric system leads to temperature rise and global warming.

Before the industrial revolution, the carbon dioxide (CO_2) in the atmosphere was about 280 parts per million. In March 2004, the atmospheric CO_2 concentration reached 377.7ppm, the largest average increase in 10 years so far. According to the statistics of scientists from the National Oceanic and Atmospheric Administration (NOAA) and the Scripps Institution of Oceanography (SIO), the monthly average concentration of CO_2 reached a peak of 421ppm in May 2022. An OECD report predicted that the CO_2 concentration would reach 685ppm by 2050.

II. Question restatement

1) Do you agree with the global temperature statement? Data given in attachment and other data sets collected by the team analyze changes in global temperature. Are you suggesting that the rise in global temperature in March 2022 has caused more growth than any decade ago. Based on historical data, please develop two or more mathematical models to show the past and predict the future level of global temperature. Use each model at 1(b) to estimate global temperature in 2050 and 2100. Which model of 1(b) do you think is the right one, and why.

2) What are the causes that affect temperature change? Use the results of question 1 and annex Data. Data from other data sets collected by csv and your team, developed a mathematical model to verify the relationship between global temperature, time and location (whatever), and explain the relationship between them or prove that there is no relationship between them. Collect native data and check actions that have an impact on natural disasters (such as volcanic eruptions, forest fires and COVID-19). Does it affect global temperature? What do you think is the main reason that is affecting global temperature change? Do you think there are steps that can stop or slow global warming?

3) Prepare a non-technical article (mostly on 1 page). Write a non-technical article (mostly on 1 page) to the APMCM Organization Committee to explain your team's findings and suggestions for the future. The PDF solution with a total number of not more than 25 pages contains: a full table of one page; and complete solution catalogue; A non-technical article page.

III. Terms and symbols

3.1 Terms

- ARIMA: autoregressive integrated moving average.
- LSTM: long and short term memory network.
- AIC: Akaike information criterion.
- SC: Schwarz criterion.
- ACF: autocorrelation function.
- cell: chain of repeated modules.

3.2 Symbols

Symbols	Meaning	Unit
y_i	real value	°C
x_i	average temperature (local or global)	°C
μ	constant	-
γ_i	coefficeint of x_i	-
θ_i	coefficeint of ϵ_i	-
$p,\ q$	lagged times	-
d	differential order	-
h_{t-1}	input before cell t-1	-
x_t	input of current cell	-
b_f	bias term	-
$\tilde{C}_t$	candidate vector	-
C_t	new position of cell t	-
$C_t - 1$	previou position of cell t	-

IV. Models assumptions

- 1. Global temperature is not affected by the outside universe (such as the sun). Only affected by the Earth itself.
- 2. Global temperature data is accurate. They can describe an aspect of global temperature.
- 3. Nature of the Earth is stable, global temperature is stable if there is no factor on the Earth.
- 4. Global temperature is affected by the factors we pick only.

V. Solution for question 1

In whole question 1, we need to analyze global temperature changes using dataset in the attachment and other datasets we collect.

5.1 Global temperature increase faster

5.1.1 *Analysis of problem 1(a)*

In this question, we need to judge if the increase of global temperature in March 2022 resulted in a larger increase thean observed over any previous 10-year period. So we can plot the global temperature in 2022. Firstly, consider if global temperature in March 2022 is a abnormal value. Then calculate the average temperature between 2011 to 2021 and its rising rate. We let them be visual and analyze their trend and differences. Finally we conclude a agreement of the statement.

5.1.2 *Solution of problem*

We download global temperature dataset from `https://berkeleyearth.org/data/`. And we choose land only monthly average temperature to plot the figure. The picture about annual average temperature between 1760 to 2022 is the following.
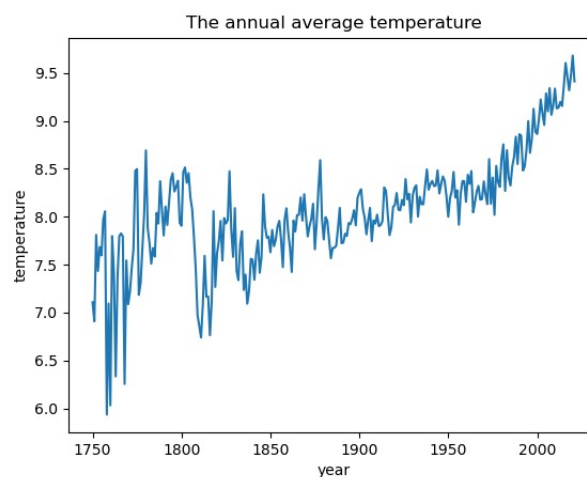

Figure 1 Annual average temperature

Then we get the average temperature in March between 2011 to 2021 is the following.

$$[6.0780, 5.9890, 6.2030, 6.3420, 6.7220, 7.4750, 7.1340, 6.6680, 7.0950, 7.0450, 6.6840].$$

The total increasing rate is 1.100%, and increase $0.606°C$ in value. Increasing rate respectively is given by

$$[-0.0941, -0.0146, 0.0357, 0.0224, 0.0599, 0.1120, -0.0456, -0.0653, 0.0640, -0.0070, -0.0512].$$

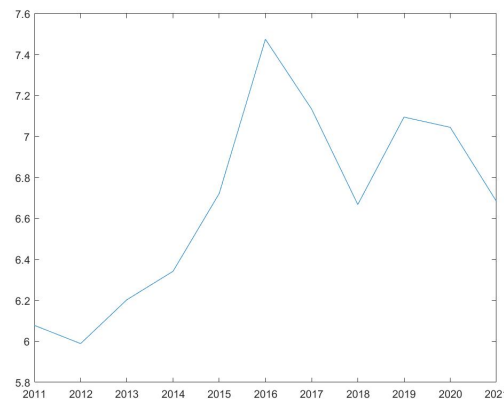Past 10 years average temperature is $6.6759°C$.

Figure 2 Near 10 years' temperature

And the monthly average temperature between January 2022 to October 2022 except March we can calculate is

$$[3.989, 4.358, 6.812, 9.665, 12.313, 14.742, 15.588, 15.07, 13.151, 10.811].$$
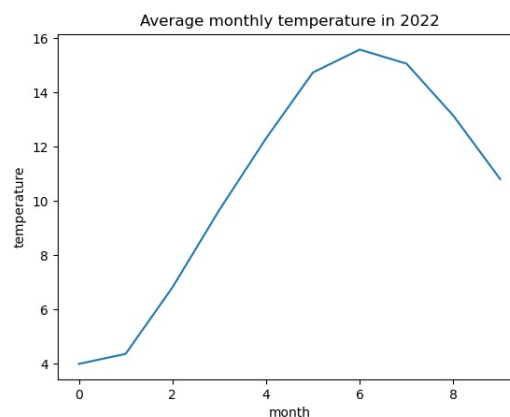

Figure 3 Average monthly temperature in 2022

fig3

Especially, the increasing rate between 2022 to 2021 is 0.0192. This is greater than previous 10 years. So we think that the increase of global temperature in March 2022 resulted in a larger increase than observed over any previous 10-year

5.2 Build model description

5.2.1 *Analysis of problem 1(b)*

Question (b) require us to build 2 or more mathematical models to describe the past and predict the future global temperature level. We can pick a observation as example to analyze the data

annually. That is we build models using yearly average temperature. We decide to use Autoregressive integrated moving average (ARIMA) model and long short term memory algorithm(LSTM) to describe the global temperature.

5.2.2 *Model 1: ARIMA*

ARIMA model, differential integrated moving average autoregression model, also known as integrated moving average autoregression model, is one of the time series prediction and analysis methods. In ARIMA (p, d, q), AR is 'autoregressive', and p is the number of autoregressive terms; MA is the 'moving average', q is the number of moving average terms, and d is the number of differences (orders) made to make it a stationary sequence. This model is built to describe the relationship between current value and historical value using the historical time data of the variable itself to predict itself. So it fit to the question in 1(b).

Firstly, we need to determine the order p which means the lagged order i.e how many previous variable related to the current value. This leads an autoregressive model(AR):

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t$$

where y_t is current value, μ is constant.

Then we consider more about the error along the time series. We get the moving average model(MA):

$$y_t = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t_i} + \epsilon_t$$

This can reduce the random fluctuation when estimating.

Finally, we combine previous 2 models then obtain the model ARMA

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t_i} + \epsilon_t$$

But this model has some limitations. The results that ARMA obtains might not be stationary, then some test will not be credible. So we can add a difference order d to let them be stationary.

We load the dataset 'land year std data'

Before we obtain the expression of ARIMA, we need to determine the value of p, q and d. In general, we choose low value of d i.e low order for the model. So it is important to find a reasonable value of p, q. We can check the Akaike information criterion(AIC) and Schwarz criterion(SC) given by respectively:

$$AIC = \ln(\frac{SSE}{T}) + \frac{2K}{T}, \quad K = p + q + 2.$$
$$SC = \ln(\frac{SSE}{T}) + \frac{K \ln(T)}{T}.$$

If AIC and SC are small enough, then the choices of p and q are preferred. So we can use the force algorithm to find the best combination of p, q.

Moreover we need to check which model we should use to estimate the data. We can check the autocorrelation function(ACF):

$$ACF(k) = \rho_k = \frac{Cov(y_t, y_{t-k})}{Var(y_t)}.$$

k is the lagged index.

We calculate the statistic ACF and plot a scatter so that we can really determine which model is the better one.

After we get the annual average temperature, we first perform a stationarity check on the data. The most intuitive way is to use the data to make a sequence diagram and then observe the trend of the image. At the same time, we calculate the first-order difference and the second-order difference of the data and make a graph to observe the trend.
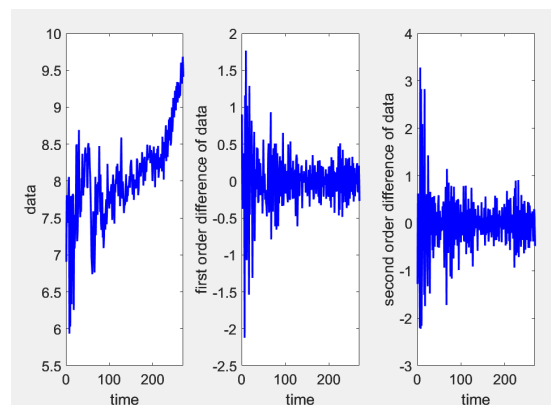

Figure 4 Different order of avg temperature

It is not obvious from the image to confirm the stationary relationship of the data, so we use the mathematical test method——ADF test and KPSS test.

In the ADF test, if the return value is 0, the data is not stationary, and if the return value is 1, the data is stationary. Then in KPSS test, if the return value is 0, the data is stationary, and if the return value is 1, the data is not stationary.

From the test result we can know the raw data is not stationary, so we need to calculate the first order difference of data, then we do the same test to the first order difference of data. We can get

```
d_adf =

   logical

    0

d_kpss =

   logical

    1
```

```
d1_adf =

   logical

    1

d1_kpss =

   logical

    0
```

Figure 5 picture 5   **Figure 6 picture 6**

Therefore, the data after the first order difference can be used for time series modeling analysis. And then what we need to do is to confirm the parameters (p, q, d). The first step is to examine the autocorrelation and partial autocorrelation plots of a stationary time series.

We can't get the values of p and q from the figure, so we adopt the AIC and BIC criteria to find the most suitable p and q values within a certain range. When we have confirmed the appropriate (p, q, d) value, we can build a model to predict the subsequent data. In the process of forecasting, we also set the upper and lower forecast limits under the 95% confidence interval.

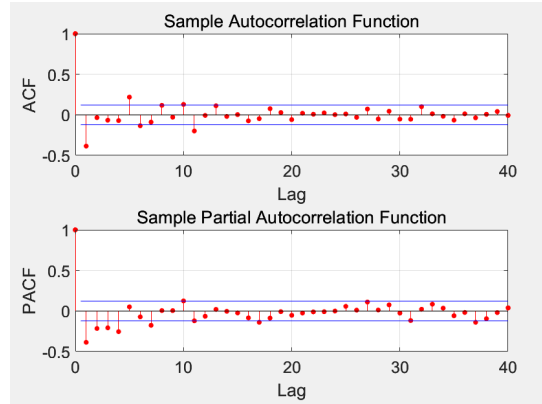Finally we get that the best combination of p, q is:

$$p = 3, \quad q = 2$$

Figure 7 ACF and PACF

5.2.3 *Model 2: LSTM*

The recurrent neural network has memory, parameter sharing and Turing completeness, so it has certain accuracy and advantages to select the recurrent neural network for learning and prediction.

LSTM, a short - and long-term memory network, is used in this paper. LSTM is formed by a chain of repeated modules, each of which is called a cell (cell).

Each cell has a specific gate structure to achieve selective information transmission. Through the transmission of LSTM gate structure (forgetting gate, input gate, cell state update, output gate) information, each cell state is updated.

The function of Amnesia Gate is to selectively update the cell state. The first step of LSTM is to determine the cell status update.

It uses the Sigmoid function to h_{t-1} and x_t processing, outputting numbers between 0-1. 1 means completely reserved, while 0 means completely forgotten. Where h_{t-1} is the weight matrix, b_f is the offset term.

$$f_t = \sigma(W_f \cdot [h_{t-1}, \ x_t] + b_f)$$

The function of the input gate is to input data and determine the information stored in the cell state. This part is divided into two steps. First, the Sigmaid layer determines the weight of the input information in the update, that is, which information needs to be updated. Next, build a candidate vector $\tilde{C}_t$ through tanh calculation at this time

$$i_t = \sigma(W_f \cdot [h_{t-1}, \ x_t] + b_f)$$
$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, \ x_t] + b_c)$$

Next update the previous cell status C_{t-1}, add this matrix and the last cell state after the forgetting gate processing to form C_t. Multiply the previous state matrix by f_t, representing selective

memory of the last cell state. Then we add $i_t \cdot \tilde{C}_t$ to the value we get , the new cell state C_t is obtained.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

Finally, the output gate will output based on the current cell state, passing the current input data through the Sigmaid layer. Then pass the current cell state through the tanh layer, normalize the matrix between -1 and 1, and perform the basic product operation with the output of the sigmoid layer, So far, the data of the output part has been determined.

$$o_t = \sigma(W_f \cdot [h_{t-1}, \ x_t] + b_o)$$
$$h_t = o_t \cdot tanh(C_t)$$

Through the module chain network composed of LSTM cell units, Be able to predict the future through continuous learning of input data.

We set the train set, test set as 2:8. And predict future 2000 years global temperature.

5.3 Prediction of global temperature

According to the model in question 1(b), we can easily predict global temperatures in 2050 and 2100 respectively using different models. Moreover we need to find when the average temperature of observation points in ous prediction models will reach 20.00°C .

5.3.1 *Prediction from model ARIMA*

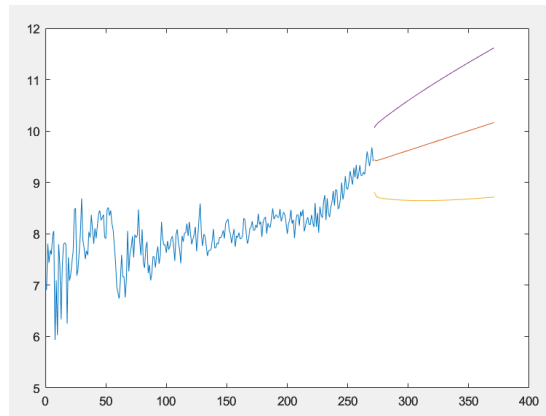First, let's predict the annual average temperature for the next 100 years.

Figure 8 Prediciton using ARIMA

From the predicted results, when it comes to 2050, the annual average temperature will reach 9.6147 °C . when it comes to 2100, the annual average temperature will reach 9.9985 °C .

Since 20 °C has not been reached, we continue to extend the forecast range. We will predict until 1400 years later. We found that when the time comes to 3404 years , the annual average temperature in the world will reach 20.0075 °C and exceed 20 °C for the first time.
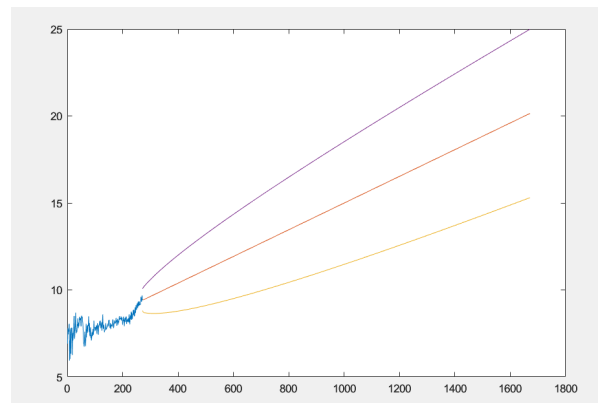
Figure 9 Further prediction using ARIMA

5.3.2 *Prediciton from model LSTM*

We enter annual standard data into the model, then we predict future 2000 years global temperature. We draw a picture about the prediction.
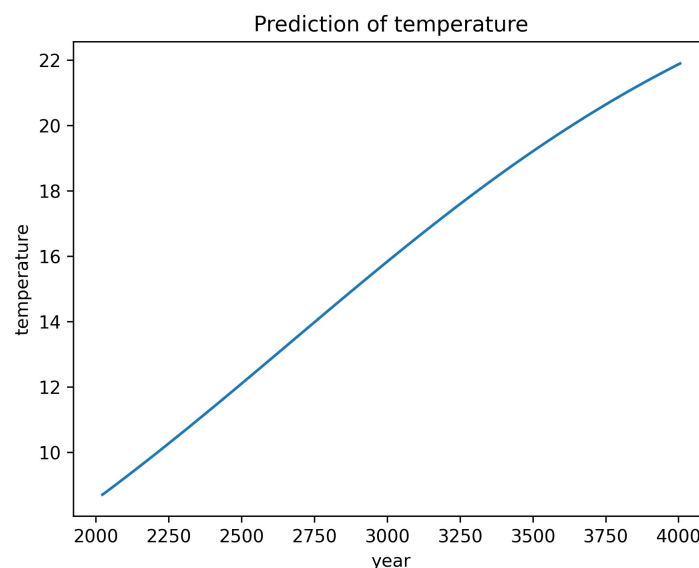
Figure 10 LSTM prediction

We can see that, annual global temperature will increase as a line in the future. We further calculate that in 2050, global temperature will be . In 2100, global temperature will be . And finally arrive to 20°C in 3633.

5.4 ARIMA is more accurate

We need to calculate important statistics to determine which model is better. We compare SE, MSE and RMSE and draw a figure as the following.

5.4.1 *Analysis of model ARIMA*

Using method of statistics, we can get the following analysis of ARIMA(3,2,1).

	Value	StandardError	TStatistic	PValue
Constant	0.0083569	0.0052983	1.5773	0.11473
AR1	-0.45471	0.15899	-2.86	0.0042368
AR2	0.27647	0.079342	3.4845	0.00049308
AR3	0.089483	0.0525	1.7044	0.0883
MA1	-0.14389	0.15508	-0.92784	0.35349
MA2	-0.63333	0.14147	-4.4768	7.5784e-06
Variance	0.10201	0.0054319	18.78	1.1057e-78

We can find that, most of the p-values are very small. So we can conclude that most of the coefficients are reliable except MA1.

5.4.2 Analysis of model LSTM

In this model LSTM, we calculate the model statistic RMSE. That is we calculate RMSE as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - f(x_i))^2}$$

where $N = 272$, y_i is the real global temperature from 1750 to 2022, and $f(x_i)$ is the prediction of it. We get $RMSE = 1.792$. So we claim that the prediction using LSTM has less reliability than model ARIMA.

In conclusion, we finally choose ARIMA to modify the description of past and future global temperature level. Because we find that ARIMA statistic is more reliable than LSTM statistic.

VI. Solution for question 2

In question 2, we are required to build a global temperature model not only including the value of temperature but also more other reasons such as position, time, volcanic eruptions, forest fires and the COVID-19 etc.

6.1 Relation about time and location

6.1.1 Analysis of problem 2(a)

According to the data given in attachment 1, we analyze the correlation between global temperature, time and location. We can use multiple regression to describe the relationship between global temperature, time and location.

6.1.2 Relationship about location

We first analyze the relationship between latitude and temperature. And we conclude that only latitude effects temperature by geometry. Since we only consider the relationship between latitude and temperature, so we can arbitary choose data in one year. We use different observation points to

denote different latitude. Then we can get average temperature in 2010, for example, while every temperature represents different latitude. And we draw a heat picture to describe the relationship.
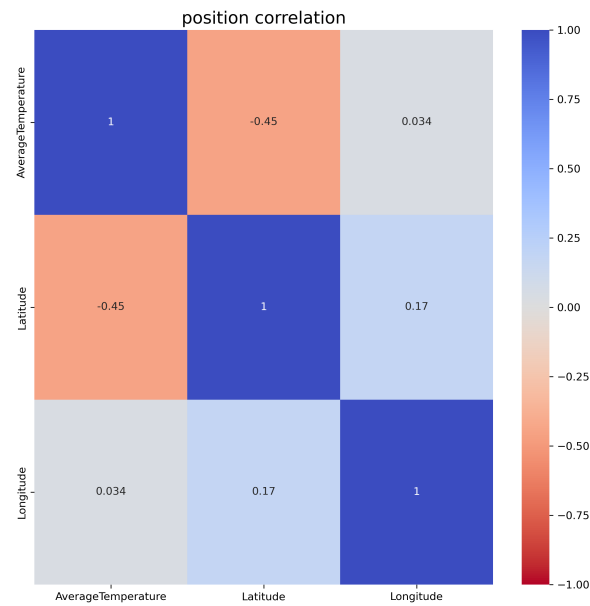

Figure 11 position correlation

Moreover, we make a statistic table for regression model.

Figure 12 regression table

We can see that p-value is small enough, so we can conclude that this model is good. And we conclude the relationship between temperature and location:

$$Avgtemperature = -0.1565 \cdot latitude + 0.0131 \cdot longitude + 22.4571.$$

6.1.3 *Relationship about time*

Then, we analyze the relationship between global average temperature and time. We choose one city Shanghai to analyze since we need to reduce the effect of location. So we pick several years

monthly average temperature of Shanghai. And we divide every year as a period to use the seasonal statsmodel package.
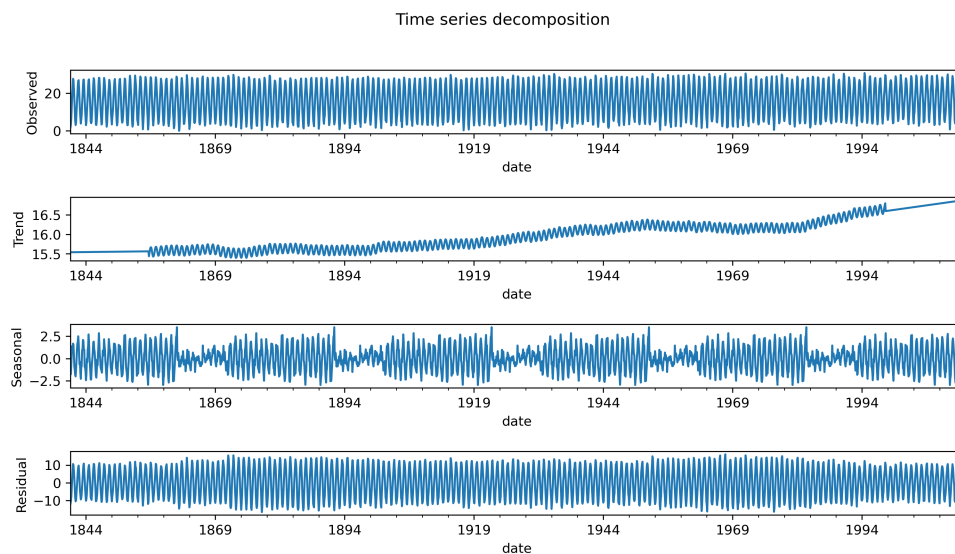
We can get the time series decomposition.

Figure 13 time series decomposition

We find that the estimation and resident are almost stable, and the observed temperature data has very small increasing trend. We need to confirm that time has no relationship to global temperature.

Then we use augmented Dickey-Fuller test to judge if time factor is a noise for average temperature. We get that the p-value is 3.421929753273345e-10. So we claim that time has no relationship to global temperature.

6.2 Disasters factors

6.2.1 *Relationship analysis*

In this section, we choose 2010 volcanic eruptions in Ayafara, Iceland to analyze the average temperature change of Iceland in 2010. We plot the average temperature of Iceland from January 2009 to December 2011.

Eyjafjallajokull erupted on 20th March, 2010 first. Then it erupted on 14th April again. We want to discover if the eruption affected the local temperature.

Let's focus on the figure. We find that the avg temperature in 2010 is different to 2009 and 2011, which has abnormal value in February and March. The temperature in this two months was on the opposite. In March, 2010, the temperature increase to $1.873°C$, while $-0.079°C$ and -0.228 in 2009 and 2011 respectively. Moreover, in April, 2010, the temperature was still a abnormal value which was $-0.208°C$. After that, there was a high avg temperature during May to October up to $1.577°C$. It is $0.677°C$ greater than which in 2009, and $1.0985°C$ greater than which in 2011. And the avg temperature $1.577°C$ is almost $0.5°C$ greater than the annual temperature of Iceland. It is a very abnormal value for Iceland. And we claim that there is a positive relationship between volcanic eruptions and global temperature.
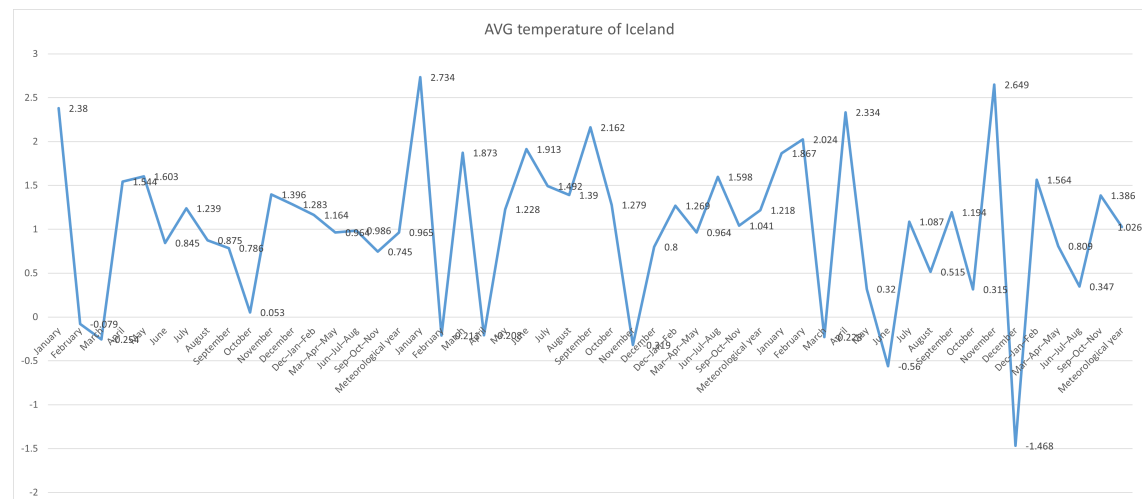
Figure 14 temperature of Iceland

Next, we analyze the relationship between forest and global temperature. We choose the 2019 Australian forest fires. The forest fires in 2019 is a major natural disaster. And we plot the average temperature of Australia from January 2018 to December 2020.
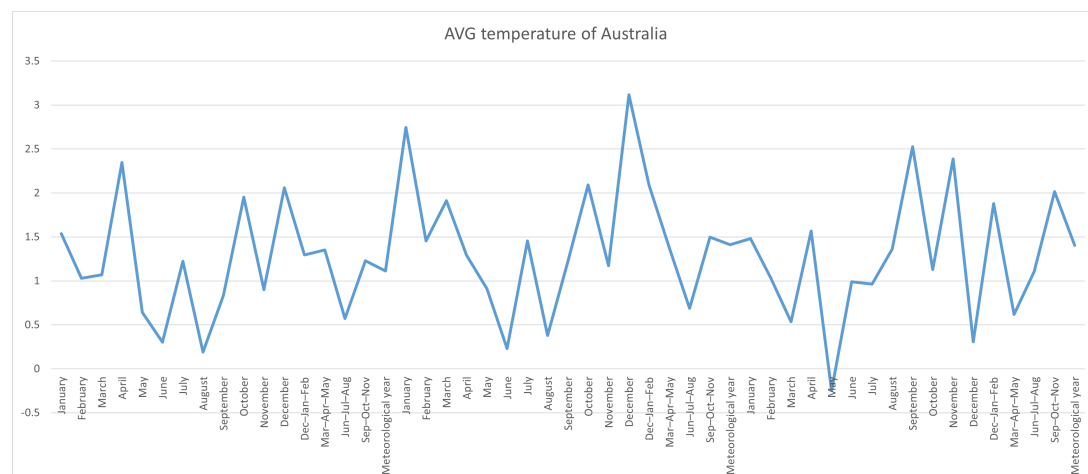
Figure 15 temperature of Australia

The Australian forest fires was last for 1 year. So we focus on July, 2019 to July, 2020.

The avg temperature during the forest fire is $1.268°C$. And there is $1.093°C$ during July, 2018 to July, 2019. It is $0.175°C$ greater. We claim that there still a positive relationship between forest fires and global temperature.

Last but not least, we consider the relationship between COVID-19 to global temperature.

We can draw the picture about injected number and temperature from 2020 Jan.

We calculate the average temperature of first 15 months is $avg1 = 9.0693$, and $avg2 = 10.5949$ represents last 31 months when COVID-19 burst. We can see that after COVID-19 burst in 2020 Jan, the global average temperature arised about $1.5256°C$. And this rise is reasonable because people need more activities to fight against COVID-19. So people around the world will take action such as large transport, factory produce etc. In other words, this will lead to entropy increase and be characterized by global average temperature rise. So we still claim that there is a positive

Figure 16 COVID-19 and temperature

relationship between global temperature and COVID-19.

6.2.2 *Summary*

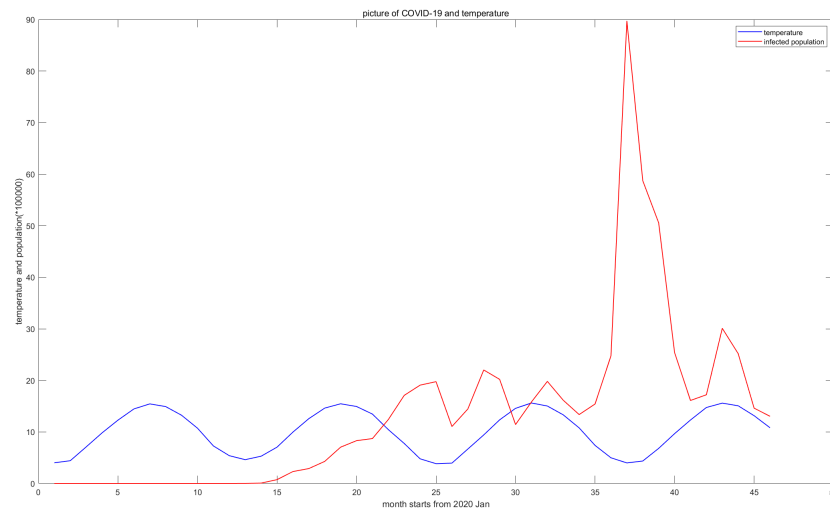In question 2(a), we search the relationship betweennatural disasters (such as volcanic eruptions, forest fires and the COVID-19) and global temperature. And take the example of 2010 Eyjafjallajokull eruption, 2019 Australian forest fires to consider the relationship about local temperature. Then expand the local conclusion to global conclusion that there is a positive relationship between volcanic eruptions, forest fires and global temperature. Moreover, we find that COVID-19 also has a positive relationship to global temperature.

Human activities fighting against this kind of disasters is a vital reason for global warming. And disasters can also be considered as a entropy increase of the Earth. They are the global activities that affecting human being. So people must produce more heat during disasters than other time. We conclude that disasters make global temperature increase.

6.3 More reason analysis based on GRA

6.3.1 *Data of more factors*

We find datasets of world energy, world population and Global forest resource.
```
https://population.un.org/wpp/Download/Standard/MostUsed/
https://fra-data.fao.org/WO/fra2020/dataDownload/
https://www.gddat.cn/newGlobalWeb/#/FeatureAnalysis
https://www.fao.org/faostat/en/#data.
```

6.3.2 *Analysis of relationship*

Then we can use Grey Relation Analysis (GRA) to estimate the relationship between other factors

6.3.3 *Build GRA*

In fact, the GRA algorithm provides a way to measure the distance between two vectors. For a time factor, a vector can be considered a time curve, while the GRA algorithm is measured if the shape and trend of the two curves are the same. In order to avoid other interference and the effect of protruding morphological features, the GRA first makes normalization, fix all vectors in the same size and position, and then calculates the distance of each point. Finally, by adjusting min function and max function the last output result falls between 0 and 1, corresponding to the general definition of * * coefficient **. Rho fixes the difference between different coefficients of correlation, in other words, the difference in output, t o be harder or more compact. From the mathematical point of view, the algorithm measures the sum of the reciprocal l_1 norm distance of each dimension of the normalized sub vector and the parent vector, and maps it to the 0 1 interval as a strategy to measure the correlation between the sub vector and the parent vector.

6.3.4 *Relationship based on GRA*

Using GRA method, we can get the result is:
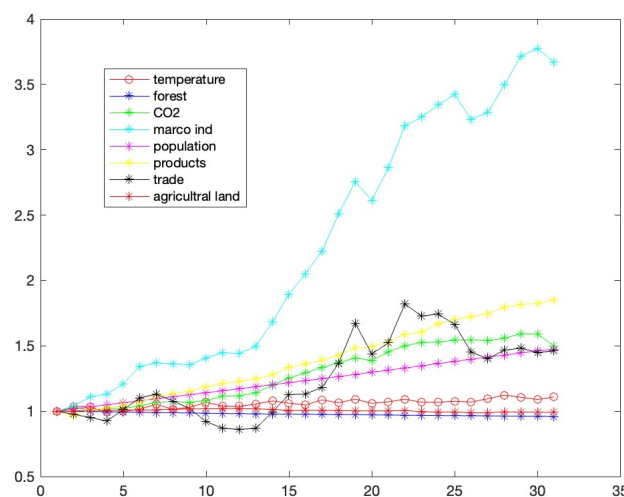
Figure 17 Normal data of factors

By using the additional 7 types of data collected, we use GRA(Grey Relation Analysis) to establish an evaluation system to determine which indicators have the greatest impact on temperature. Serialize and normalize the 7 types of data, and then use the established formula to calculate the corresponding gray correlation coefficient and the average value of the correlation coefficient to obtain the degree of influence of different indicators on temperature. This method reflects the association degree between two sequences by measuring the distance between two normalized vector sequences and constraining it between 0 and 1. We can get the result below

Index	forest	CO2	macro ind	population	products	trade	agricultural
Correlation	0.9437	0.8653	0.6097	0.8871	0.8177	0.8538	0.9602

Figure 18 Normal data of factors

From the values in the table, it can be seen that agricultural area, forest area and population factors have a greater impact on temperature.

This is obvious. Population growth is one of the main factors contributing to global warming. At the same time, it seriously threatens the balance between the natural ecological environment. Excessive population leads to the continuous increase of CO2 content 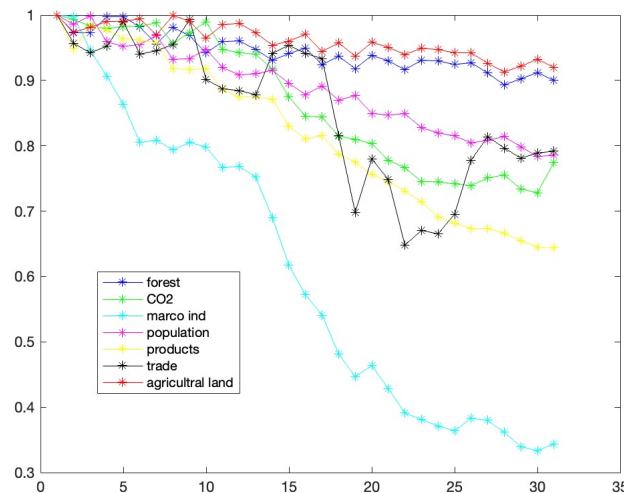in the atmosphere, and the resulting greenhouse effect will directly affect the climate change on the earth's surface. Along with the increase in CO2 in the air, forests play an important role in converting CO2. But forests are gradually shrinking and the amount of CO2 converted is also decreasing. Changes in agricultural area will cause changes in the absorption of greenhouse gas emissions, ground reflectivity, and evaporation, thereby causing changes in the carbon storage capacity, energy balance, and water transport of the entire ecosystem. The urban heat island effect is the best example of the impact of residential expansion on the local climate, and microclimate changes caused by land cover changes also occur in life. Through the interaction between various scales and various spheres on the earth's surface, the impact of land use on the climate on a small local scale can be transferred to a larger scale or even a global scale, and the urban heat island effect will have great influence on the increase of global temperature.

6.4 Measures to curb or slow down global warming

In question 2(d), we need to find some measures to curb or slow down global wariming. We think there are 3 measures to curb global warming. The first is to control population growth. Effective control of population growth reduces CO2 emissions. The second is to plant trees, so that forests can play a greater role in absorbing and transforming CO2. The third is to improve the utilization technology and utilization rate of various energy sources, make more efficient use, and reduce CO2 emissions.

VII. A non-technical article

Global warming is a phenomenon related to nature. Due to the continuous accumulation of the greenhouse effect, the energy absorbed and emitted by the earth-atmosphere system is unbalanced, and the energy is continuously accumulated in the earth-atmosphere system, which leads to an increase in temperature and global warming. When people burn fossil fuels, such as oil, coal, etc., or cut down forests and burn them, a large amount of carbon dioxide, or greenhouse gas, is produced. These greenhouse gases are highly transparent to visible light from solar radiation and emitted to the earth. The long-wave radiation is highly absorptive, which can strongly absorb the infrared rays in the ground radiation, leading to an increase in the temperature of the earth, that is, the greenhouse effect. Global warming will redistribute global precipitation, melt glaciers and permafrost, and rise sea levels, which will not only endanger the balance of natural ecosystems, but also affect human health and even threaten human survival. On the other hand, due to terrestrial greenhouse gas emissions, the temperature of the continent has risen, and the temperature difference between the ocean and the ocean has become smaller, which in turn has caused the air flow to slow down, and the smog cannot be blown away in a short time, resulting in more smog in many cities and affecting human health. Measures such as car restrictions and suspension of production have only short-term and partial effects, and cannot fundamentally change climate warming and smog pollution.

The previous part of this article shows that the global average temperature is showing a rising trend year by year, especially since humans entered the industrialization period, the global average temperature has risen more significantly. After research and prediction, it is found that the global average temperature will approach 10 degrees Celsius in 2100, and in the next 1,500 years, the global average temperature may reach 20 degrees Celsius.

Studies have shown that the global average temperature rise is affected by many aspects, including forest area, CO2 emissions, macroeconomic indicators, population, total production, total trade, agricultural land area, etc. Among these factors, the agricultural land area has a significant impact on the global average temperature rise Higher impacts are greatest, and macroeconomic indicators have the smallest impact on global average temperature rise.

Knowing the conclusion above, we human being must take action to slow down the global temperature increasing. Firstly, we must control the population growth. Although we cannot force people not to give birth of children, we can show the awareness of Oligogenesis and superior breeding. Then the increase rate of population could be slowed down. Moreover, since the forest land and agricultral land own vital relationship to global temperature, we plant more tree and recover more green land are vital action. If we lessen the cut of the tree and develop utilization technology, human will release less CO2. Then the global temperature increasing will slow down.

Data Source:

- Global Disaster Data Platform: `https://www.gddat.cn/newGlobalWeb/#/FeatureAnalysis`
- FAO data page: `https://www.fao.org/faostat/en/#data`
- WHO Coronavirus data page: `https://covid19.who.int/data`

VIII. References

[1] FAN Bihang, Li Ning, GUO Li, Relationship between the Number of Natural Disasters and Climate Change from 1950 to 2010 in China, Academy of Disaster Reduction and Emergency Management,Ministry of Civil Affairs & Ministry of Education, Beijing 100875.

[2] 石头, 时间序列 (ARIMA) 模型及其 matlab 实现, `https://zhuanlan.zhihu.com/p/495845415`, 2022, 11, 25.

[3] : FAO. 2020. Global Forest Resources Assessment 2020 –Key findings. Rome. `https://doi.org/10.4060/ca8753en`

IX. Appendix

Listing 1: The matlab source code of ARIMA

```matlab
clear all
clc
date=1751:2021;
rawdata=xlsread('land year std data.xls');
data=rawdata';
%rawdata=xlsread('land avg monthly.xlsx');
%data=rawdata([1800:3274],13)';
subplot(131)
plot(data,'b-','LineWidth',2)
xlabel('time')
ylabel('data')
set(gca,'fontsize',15)
subplot(132)
ddata = diff(data);
plot(ddata,'b-','LineWidth',2)
xlabel('time')
ylabel('first order difference of data')
set(gca,'fontsize',15)
subplot(133)
dddata = diff(data,2);
plot(dddata,'b-','LineWidth',2)
xlabel('time')
ylabel('second order difference of data')

set(gca,'fontsize',15)


% Stationarity Test of the raw data
d_adf=adftest(data)
d_kpss=kpsstest(data)

%if adftest=0 or kpsstest=1 that means the data is not stationary
ddata = diff(data);
d1_adf = adftest(ddata)
d1_kpss = kpsstest(ddata)

%if adftest=1 or kpsstest=0 that means the data is stationary
%ddata=data;

% draw the ACF graph and PACF graph
figure
subplot(211)
autocorr(ddata,40)
ylabel('ACF')
set(gca,'fontsize',15)
```

```matlab
    subplot(212)
    parcorr(ddata,40)
    ylabel('PACF')
    set(gca,'fontsize',15)

    %% calculate the value of q and p
    pmax = 4;
    qmax = 4;
    d = 1;
    [p, q ]=findPQ(data,pmax,qmax,d);

    %% bulid model
    %p = 3;q = 2;
    Mdl = arima(p, 1, q); %first order difference
    EstMdl = estimate(Mdl,data');
    %% predict model
    step = 2000;
    [forData,YMSE] = forecast(EstMdl,step,'Y0',data');
    lower = forData - 1.96*sqrt(YMSE); %the lower bound of the 95% confident interval
    upper = forData + 1.96*sqrt(YMSE); %the upper bound of the 95% confident interval

    figure
    plot(1:length(data),data)
    hold on
    plot((length(data)+1):length(data)+step,forData)
    hold on
    plot((length(data)+1):length(data)+step,lower)
    plot((length(data)+1):length(data)+step,upper)

    function [p q] = findPQ(data,pmax,qmax,d)
    data = reshape(data,length(data),1);
    LOGL=zeros(pmax+1,qmax+1);
    PQ=zeros(pmax+1,qmax+1);
    for p=0:pmax
        for q=0:qmax
            model=arima(p,d,q);
            [fit,~,logL]=estimate(model,data); %strcture
            LOGL(p+1,q+1)=logL;
            PQ(p+1,q+1)=p+q; %number of parameters
        end
    end
    LOGL=reshape(LOGL,(pmax+1)*(qmax+1),1);
    PQ=reshape(PQ,(pmax+1)*(qmax+1),1);
    m2 = length(data);
    [aic,bic]=aicbic(LOGL,PQ+1,m2);
    aic0 = reshape(aic,(pmax+1),(qmax+1));
    bic0= reshape(bic,(pmax+1),(qmax+1));

    aic1 = min(aic0(:));
```

```matlab
index = aic1==aic0;
[pp qq] = meshgrid(0:pmax,0:qmax);
p0 = pp(index);
q0 = qq(index);

aic2 = min(bic0(:));
index = aic2==bic0;
[pp qq] = meshgrid(0:pmax,0:qmax);
p1 = pp(index);
q1 = qq(index);

if p0^2+q0^2> p1^2+q1^2
    p = p1;
    q = q1;
else
    p = p0;
    q = q0;
end
end
```

Listing 2: The python source code of LSTM

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from tensorflow import keras
from tensorflow.keras import layers
from sklearn.preprocessing import MinMaxScaler
from keras import regularizers

#### 数据处理部分 ####

# 读入数据
google_stock = pd.read_excel('land year std data.xlsx')
# google_stock.tail() # 查看部分数据
# google_stock.head()

# 时间戳长度
time_stamp = 10 # 输入序列长度

# 划分训练集与验证集
google_stock = google_stock[['std data']]
lt = 500
train = google_stock[0:272]
valid = google_stock[:]
test = google_stock[272:]

# 归一化
scaler = MinMaxScaler(feature_range=(0, 1))
```

```python
scaled_data = scaler.fit_transform(train)
x_train, y_train = [], []

# 训练集切片
for i in range(time_stamp, len(train) - 5):
    x_train.append(scaled_data[i - time_stamp:i])
    y_train.append(scaled_data[i: i + 5])

x_train, y_train = np.array(x_train), np.array(y_train).reshape(-1, 5)

# 验证集切片
scaled_data = scaler.fit_transform(valid)
x_valid, y_valid = [], []
for i in range(time_stamp, len(valid) - 5):
    x_valid.append(scaled_data[i - time_stamp:i])
    y_valid.append(scaled_data[i: i + 5])

x_valid, y_valid = np.array(x_valid), np.array(y_valid).reshape(-1, 5)

# 测试集切片
scaled_data = scaler.fit_transform(test)
x_test, y_test = [], []
for i in range(time_stamp, len(test) - 5):
    x_test.append(scaled_data[i - time_stamp:i])
    y_test.append(scaled_data[i: i + 5])

x_test, y_test = np.array(x_test), np.array(y_test).reshape(-1, 5)
#### 建模部分 ####
model = keras.Sequential()
model.add(layers.LSTM(64, return_sequences=True, input_shape=(x_train.shape[1:])))
model.add(layers.LSTM(64, return_sequences=True))
model.add(layers.LSTM(32))
model.add(layers.Dropout(0.1))
model.add(layers.Dense(5))

model.compile(optimizer=keras.optimizers.Adam(), loss='mae', metrics=['accuracy'])
learning_rate_reduction = keras.callbacks.ReduceLROnPlateau(monitor='val_loss',
    patience=3, factor=0.7,
                                                min_lr=0.000000005)

history = model.fit(x_train, y_train,
                    batch_size=128,
                    epochs=100,
                    validation_data=(x_valid, y_valid),
                    callbacks=[learning_rate_reduction])

# loss变化趋势可视化
# plt.plot(history.history['loss'],label='training loss')
# plt.plot(history.history['val_loss'], label='val loss')
```

```python
# plt.show()
closing_price = model.predict(x_test)
model.evaluate(x_test)
scaler.fit_transform(pd.DataFrame(valid['std data'].values))

# 反归一化
closing_price = scaler.inverse_transform(closing_price.reshape(-1, 5)[:,
    0].reshape(1, -1)) # 只取第一列
y_test = scaler.inverse_transform(y_test.reshape(-1, 5)[:, 0].reshape(1, -1))

# 计算预测结果
rms = np.sqrt(np.mean(np.power((y_test[0:1, 5:] - closing_price[0:1, 5:]), 2)))
print(rms)
print(closing_price.shape)
print(y_test.shape)

# 预测效果可视化
dict_data = {
    'Predictions': closing_price.reshape(1, -1)[0],
    'actual': y_test[0]
}


year =[]
for i in range(2022, 2022+1984):
    year.append(i)

data_pd = pd.DataFrame(dict_data)
plt.plot(year, data_pd[['Predictions']][0:5000])

plt.xlabel('year')
plt.ylabel('temperature')
plt.title('Prediction of temperature')
# plt.plot(data_pd[['Predictions']],linewidth=1.2)
# figure = plt.figure(figsize=(10,10))
# axes = figure.add_subplot(111) #插入面板2
# axes.plot(data_pd[0:30])
# axes.set_xlabel('tick')
# axes.set_ylabel('price')
# axes.set_title('30tick')
# plt.show()
# plt.plot(data_pd[0:30])
dataframe = pd.DataFrame(data_pd[0:2000])
dataframe.to_excel('Q1.xlsx')
plt.savefig('./q1b', dpi=320)
```

Listing 3: The python source code of multiple analysis

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import statsmodels.api as sm
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.stattools import adfuller

data = pd.read_csv('2022_APMCM_C_Data.csv',encoding='gb18030')
data.head(10)

data['dt'] = pd.to_datetime(data['dt'])
data.index = data['dt']
data['Year'] = data['dt'].dt.year
data_2010 =
    data[data['Year']==2010][['AverageTemperature','City','Latitude','Longitude']]

def clear_Latitude(s):
    if 'N' in s:
        return float(s.replace('N',''))
    else:
        return -float(s.replace('S',''))
def clear_Longitude(s):
    if 'E' in s:
        return float(s.replace('E',''))
    else:
        return -float(s.replace('W',''))
data_2010['Latitude'] = data_2010['Latitude'].apply(clear_Latitude)
data_2010['Longitude'] = data_2010['Longitude'].apply(clear_Longitude)

city_data_df = data_2010.groupby('City').mean()

f, ax = plt.subplots(nrows=1, ncols=1, figsize=(8, 8))
df = city_data_df
shifted_cols = df.columns
corrmat = df[shifted_cols].corr()

sns.heatmap(corrmat, annot=True, vmin=-1, vmax=1, cmap='coolwarm_r')
ax.set_title('position correlation', fontsize=16)
plt.tight_layout()
plt.savefig('fig/position correlation.png',dpi=300, bbox_inches = 'tight')
plt.show()

Y= city_data_df['AverageTemperature']
X1 = city_data_df.drop(columns=['AverageTemperature'])
X= sm.add_constant(X1)
result = sm.OLS(Y,X).fit()
result.summary()
```

```python
print(result.summary())

China_data = pd.read_csv('clear_China_city_AvgTemperature.csv')
China_data.head(10)
China_data.index =pd.to_datetime(China_data['date'])
Shanghai_data = China_data[China_data['City'] == 'Shanghai']
Shanghai_data['AverageTemperature'][Shanghai_data['AverageTemperature']<-18]=np.nan

Shanghai_data = Shanghai_data.fillna(method = 'ffill',axis=0)
Shanghai_data.AverageTemperature.plot(figsize=(20, 5))
plt.xlabel('Average Temperature')
plt.title('Average Temperature in Shanghai')
plt.savefig('fig/Shanghai historical temperature .png',dpi=300, bbox_inches = 'tight')
plt.show()

decomposition = seasonal_decompose(Shanghai_data['AverageTemperature'],
    model='additive', period=365, extrapolate_trend='freq')
fig, axs = plt.subplots(4, figsize=(12, 6))
fig.suptitle('Time series decomposition')
decomposition.observed.plot(ax=axs[0])
decomposition.trend.plot(ax=axs[1])
decomposition.seasonal.plot(ax=axs[2])
decomposition.resid.plot(ax=axs[3])
axs[0].set_ylabel('Observed')
axs[1].set_ylabel('Trend')
axs[2].set_ylabel('Seasonal')
axs[3].set_ylabel('Residual')
plt.subplots_adjust( hspace=1 )
plt.savefig('fig/seasonal.png',dpi=300, bbox_inches = 'tight')

adf_test = adfuller(decomposition.resid)
print(f"p-value = {adf_test[1]}")
```

Listing 4: The matlab source code of GRA

```matlab
% Grey relation analysis
clear all;close all;clc

zongshouru = [3439, 4002, 4519, 4995, 5566];
daxuesheng = [341, 409, 556, 719, 903];
congyerenyuan = [183, 196, 564, 598, 613];
xingjifandian = [3248, 3856, 6029, 7358, 8880];

filename='./total (c).xlsx';
dt=xlsread(filename,'Sheet2','A2:I32');

yr=reshape(dt(:,1),1,31)
fr=reshape(dt(:,2),1,31)
c2=reshape(dt(:,3),1,31)
```

```matlab
mi=reshape(dt(:,4),1,31)
pl=reshape(dt(:,5),1,31)
pd=reshape(dt(:,6),1,31)
td=reshape(dt(:,7),1,31)
al=reshape(dt(:,8),1,31)
tp=reshape(dt(:,9),1,31)

x0=tp
x1=fr
x2=c2
x3=mi
x4=pl
x5=pd
x6=td
x7=al
% normalization
x0 = x0 ./ x0(1);
x1 = x1 ./ x1(1);
x2 = x2 ./ x2(1);
x3 = x3 ./ x3(1);
x4 = x4 ./ x4(1);
x5 = x5 ./ x5(1);
x6 = x6 ./ x6(1);
x7 = x7 ./ x7(1);


% global min and max
global_min = min(min(abs([x1; x2; x3;x4;x5;x6;x7] - repmat(x0, [7, 1]))));
global_max = max(max(abs([x1; x2; x3;x4;x5;x6;x7] - repmat(x0, [7, 1]))));
%set rho
rho = 0.5;
%calculate zeta relation coefficients
zeta_1 = (global_min + rho * global_max) ./ (abs(x0 - x1) + rho * global_max);
zeta_2 = (global_min + rho * global_max) ./ (abs(x0 - x2) + rho * global_max);
zeta_3 = (global_min + rho * global_max) ./ (abs(x0 - x3) + rho * global_max);
zeta_4 = (global_min + rho * global_max) ./ (abs(x0 - x4) + rho * global_max);
zeta_5 = (global_min + rho * global_max) ./ (abs(x0 - x5) + rho * global_max);
zeta_6 = (global_min + rho * global_max) ./ (abs(x0 - x6) + rho * global_max);
zeta_7 = (global_min + rho * global_max) ./ (abs(x0 - x7) + rho * global_max);


%show
figure;
plot(x0, 'ro-' )
hold on
plot(x1, 'b*-')
hold on
plot(x2, 'g*-')
hold on
plot(x3, 'c*-')
hold on
```

```matlab
plot(x4, 'm*-')
hold on
plot(x5, 'y*-')
hold on
plot(x6, 'k*-')
hold on
plot(x7, 'r*-')
legend('temperature', 'forest', 'CO2', 'marco ind', 'population', 'products',
    'trade', 'agricultral land')

figure;
plot(zeta_1, 'b*-' )
hold on
plot(zeta_2, 'g*-')
hold on
plot(zeta_3, 'c*-')
hold on
plot(zeta_4, 'm*-')
hold on
plot(zeta_5, 'y*-')
hold on
plot(zeta_6, 'k*-')
hold on
plot(zeta_7, 'r*-')
hold on
legend('forest', 'CO2', 'marco ind', 'population', 'products', 'trade', 'agricultral
    land')
```