# Multi-Level Modelling of Global Warming Based on Machine Learning Algorithms

Xianfei Cai[1], Lijia Liu[2], Zheng Wang[1,*]

[1]Department of Information Science and Technology, Tianjin University of Finance and Economics, Tianjin, China, 300222

[2]International Business school, Tianjin University of Finance and Economics, Tianjin, China, 300222

*Corresponding author: wangzhen@tjufe.edu.cn

**Abstract.** Today, abnormally high temperatures around the world are a wake-up call for global warming. Global warming is a problem that needs to be addressed urgently. In order to assess past temperature changes, this paper constructs indicators of data growth rates over a certain period of time. Firstly, this paper analyses the changes in growth rates after 1800. In addition, this study uses the Mann-Kendall method to determine the trend of the growth rate, and the ridge regression model, ARIMA model, and LSTM model are constructed to predict the future temperature change, respectively. The results show that the LSTM model is the most reliable, whereas the ARIMA model is only suitable for short-term forecasts due to the lack of a follow-up dataset and is less effective in predicting global temperatures below 20°C. The results of the ARIMA model are summarised as follows, with polynomial fitting results being affected by the extremes of the data, and performing the worst. The results show that the temperature will reach 20°C around 2310 and within 2398~2439 according to the ridge regression model and LSTM model, respectively.

**Keywords:** Global Warming, Mann-Kendall Method, Ridge Regression, ARIMA, LSTM.

## 1. Introduction

### 1.1. Problem Background

In recent years, Global Warming has become an inevitable trend. For 2.5 million years, Earth's climate has fluctuated cyclically from ice ages to warm periods. However, the Earth's temperature has risen unusually fast in the last century, by about 1.2 to 1.4 degrees of alternate altitude. Scientists believe that human activity is driving the temperature rise, a process known as global warming. The latest series of Intergovernmental Panel on Climate Change (IPCC) climate reports (AR5) confirm the fact of global warming in the last century from multiple dimensions and perspectives and confirm that anthropogenic impacts are the leading cause of warming since the mid-20th century.

Global warming brings about extreme weather that can exacerbate global warming itself. However, the Earth's temperature system is extremely complex, related to greenhouse gas concentrations, geological movements, and human activity. Given those problems, the 2015 Paris agreement allows individual countries to take the first step in dealing with the greenhouse effect by institutionalizing their own emissions reduction programs. Moreover, there are undoubtedly two ways to reduce emissions: to develop new and clean energy sources and to be carbon neutral. At the same time, reducing emissions also implies economic benefits.

### 1.2. Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems.

To begin with, understanding and evaluating the solution to global warming is crucial to developing the human economy and society. On this issue, it is of great significance to construct the growth rate index to evaluate past temperature changes and to establish models to predict the future temperature and its trend. Therefore, we preprocessed the past data, calculated the monthly growth

rate, and performed a trend analysis on the 120 data points to determine whether the March 2022 global temperature increase is in the past decade. Then, the three models are used to analyze and predict the temperature in 2050 and 2100 and to answer the question of when the average temperature will jump by 20℃.

## 2. Multi-level Data-driven Prediction Models for Global Average Temperature Analysis

### 2.1. Detection Model of Temperature Growth Rate based on the M-K Method

The global average surface temperature rose from 0.6 to 0.9℃ between 1906 and 2005. The rate of temperature increase has nearly doubled in the last 50 years [1]. To analyze whether the rise of global temperature in March 2022 leads to a more significant increase than that over the previous 10 years, we take the growth rate of global average temperature as an evaluation index and calculate the sharp points and intervals of the growth rate series through the M-K test to conclude.

#### 2.1.1 Data Pre-processing

We obtained the monthly average temperature from the dataset provided by the APMCM organizing committee. We need to calculate the annual average temperature from the monthly average temperature. According to the statistics, there are 11002 **missing values** in the data, accounting for about 4.82%. We consider that the temperature value should be between -60 and 50℃. So, it is an outlier. Both missing values and outliers tend to affect the calculation results. Therefore, during the **data preprocessing**, we use the monthly average temperature that occurs most often in that year to replace the missing data. At the same time, we abandon the outliers. In this way, the impact of missing values and outliers on the accuracy of model data is weakened. Then, we take the average temperature of 12 months in a year as the annual global average temperature, and draw the global average temperature trend, as shown in Figure 1.

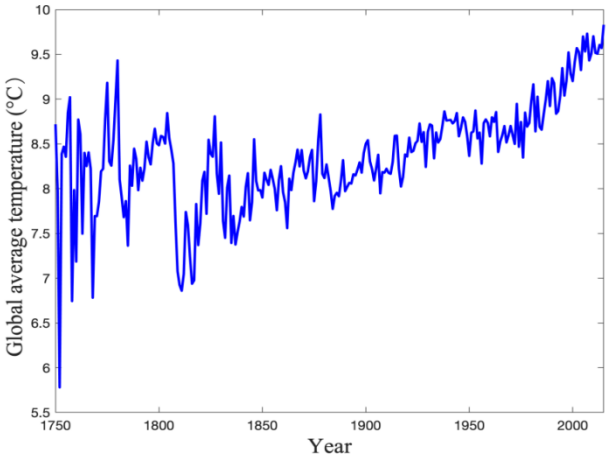



**Figure 1.** Global Average Temperature Trend.

According to Figure. 1, the annual average temperature is approximately between 5.7℃ and 10℃ from 1750 to 2022 with an overall rising trend. Due to the constant fluctuation, it is impossible to judge whether the global temperature increase in March 2022 has resulted in a more significant increase than observed over any previous 10-year period based on the trend of annual average temperature alone. Therefore, we will then introduce the concept of the growth rate.

#### 2.1.2 The Calculation of Growth Rate

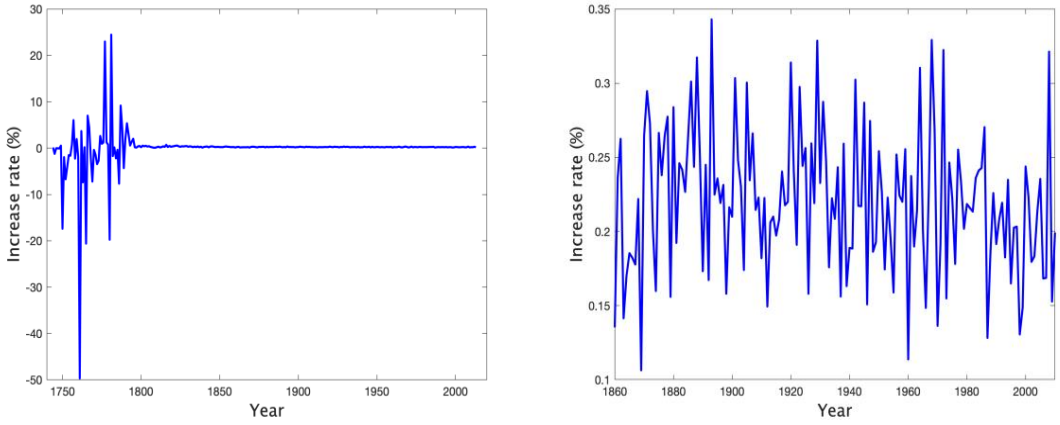
The growth rate refers to the ratio between the growing amount of data metrics in a certain period and the base data. In this model, we need to determine whether the global temperature increase in March 2022 has resulted in a more significant increase than observed over the previous 10-year period. It means that the data metrics currently are the average temperature in March 2022. So, we count the

annual growth rate of the average temperature in March over 10 years. It is necessary to take the average temperature of March as the base data and the difference between the average temperature of March of that year and the average temperature of March ten years ago as the growth amount. Then calculate the ratio between them and get the annual growth rate for March. The specific calculation formula for the growth rate is as follows:

$$I(t) = \frac{T(t)-T(t-1)}{T(t-1)} \quad (1)$$

Where $I(t)$ is the growth rate in year $t$, $T(t)$ is the value of temperature in year $t$. By doing that, we can get the growth rate in March of each year from 1750 to 2022 and quickly judge the growth rate trend.

From the trend of growth rates shown in Figure 2 (a), before 1800, the growth rate fluctuated wildly and changed irregularly. On the contrary, the growth rate after 1800 fluctuated less than before 1800, almost tending to a straight line. Therefore, we can think of the annual global average temperature in March before 1800 as a set of random variables, which can be regarded as noise signals. Therefore, we only analyze the change in growth rate since 1800.




(a) since 1750 (b) since 1800

**Figure 2.** Growth Rate.

Because the question is whether the rise in global temperature in March 2022 is causing a more significant increase than over any previous 10-year period, it is necessary to determine whether the growth rate in March 2022 is the fastest. Based on the above discussion, we plot the growth rate after 1800 in Figure 2 (b). It is obvious that the figure that the growth rate after 1800 was between 0.1 and 0.35, but it was also in a state of fluctuation. The trend of growth rate could not be judged from the sense. Therefore, **our model adopted the Mann-Kendall (M-K) method** to help us judge the growth rate change.

### 2.1.3 Mann-Kendall Method

The M-K method can detect the mutation points by calculating the statistical sequences $UF_k$, $UB_k$ and analyzing their trend graphs. It is usually used on time series trend analysis, such as precipitation, streamflow and so on. M-K method not only simplifies calculation, but also makes explicit the start and the period of mutation.

In our model, to judge the changing trend of the growth rate, we need to use the M-K method to find the mutation points in the growth rate series and the mutation interval. Therefore, we first calculate the growth sequence $UF_k$ over time and the reverse growth sequence $UB_k$. Here, we give a significance level $\alpha$ equal to 0.05, and the critical value $U_\alpha$ is 1.96. We draw them on a rectangular coordinate system. The drawing results are shown in Figure 3.

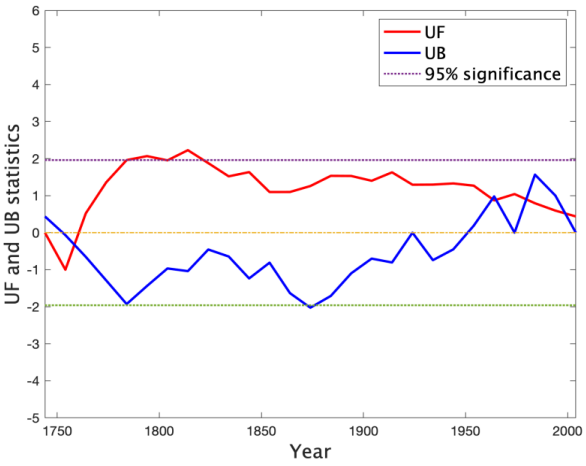



**Figure 3.** The Result of M-K Method.

As shown in Figure. 3, the red line is $UF$, and the blue line is $UB$. We can find $that\ UF$ is always bigger than zero, meaning that the growth rate since 1800 has been rising. However, it does not go beyond the line of 95% significance. In other words, there was no abrupt change in growth rates after the 1800 year. In addition, even if we consider the growth rate between 1750 and 1800, as Fig. 3(a) shows, the increase in March 1780 is the largest but not in 2022, which broke through 20%. To sum up, the global temperature increase in March 2022 does not result in a more considerable increase than observed over any previous 10-year period.

### 2.2. A Temperature Prediction Model based on Time Series

#### 2.2.1 Temperature Prediction Model Based on Ridge Regression

**(1) Description of Ridge Regression Model**

The connection of things is often multifaceted. Therefore, it is more effective and practical to predict or estimate the dependent variable by the optimal combination of multiple independent variables compared with only one independent variable. The purpose of the Multiple Linear Regression model is to construct a regression equation using multiple independent variables to estimate the dependent variable to explain and predict the value of the dependent variable. Its general form is:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n + e \tag{2}$$

Where $y$ can be approximately expressed as a linear function of the independent variables $x_i (i = 1,2,\ldots,n)$. $\alpha_0$ is the constant term. $\alpha_j (j = 1,2,\ldots,n)$ is partial regression coefficient. $e$ is the random error after removing the influence of m independent variables on $y$.

To avoid the problem of over-fitting and under-fitting, the Ridge Regression is then utilized. Ridge Regression is a modified least-squares estimation method with 2-norm regularization. It can obtain a more stable and reliable regression coefficient by losing part of the information and reducing accuracy at an unbiased cost. This method is better than the least square method for ill-conditioned data fitting. In this method, we introduce a penalty term to avoid overfitting or underfit the model. The calculation formula is as follows:

$$T_{pred} = \sum_i \omega_i t^i + \lambda \left\| t^i \right\|_2^2 \tag{3}$$

Where $T_{pred}$ is the value of temperature, we predicted. $\omega_i$ is the partial regression coefficient. $t$ is normalized by mean and standard deviation. $\lambda$ is the penalty coefficient. We can predict the temperature in the future by Ridge Regression.

### (2) The Process of Ridge Regression Model

Starting from the standard multivariate linear regression model, we studied the relationship between temperature and year. To make the calculation more accessible and eliminate the influence of dimensionality and the various size of the variable itself, we first standardized the year. The mean value is 1880, and the standard deviation is 76.93. After repeated testing, we finally adopted the multivariate form as follows:

$$y = \omega_1 x^3 + \omega_2 x^2 + \omega_3 x + \omega_4 \tag{4}$$

We use the global mean temperature from 1750 to 2022 as training data for model fitting. We obtained that the value of $\omega_1$ is between -0.04195 and 0.05444, the value of $\omega_2$ is between 0.2524 and 0.3369, the value of $\omega_3$ is between 0.1978 and 0.3864, the value of $\omega_4$ is between 8.07 and 8.183. So, we chose the best value 0.006249, 0.2946, 0.2921, 8.126.

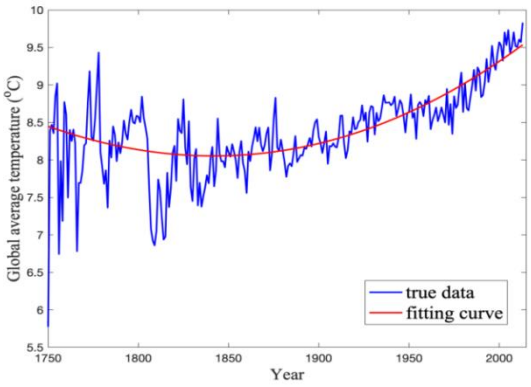



**Figure 4.** The Result of Ridge Regression

In Figure 4, the blue line is actual data, and the red line is a fitting curve. The trend of the fitting curve is almost consistent with actual data. It indicates that our model has an excellent fitting effect.

### 2.2.2 Temperature Prediction Model based on ARIMA

### (1) Description of ARIMA Model

Autoregressive Integrated Moving Average (ARIMA) tries to separate the signal from the noise and then extrapolates the signal into the future to get predictions [2]. It is highly effective for fitting non-stationary data. The structure of ARIMA is shown in Figure 5.

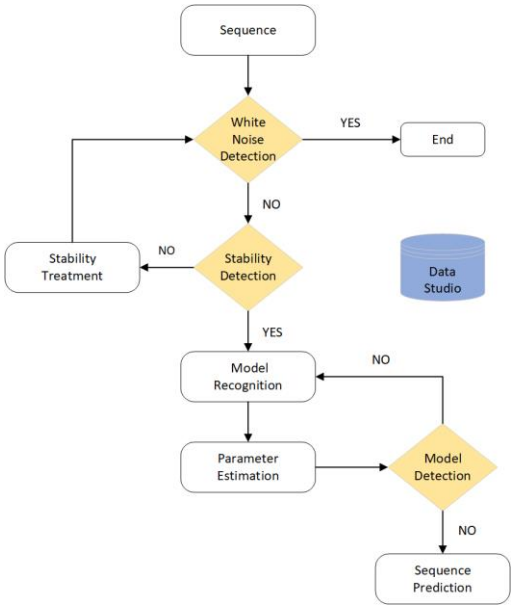



**Figure 5.** ARIMA Structure.

In ARMIA model, we first need to conduct a stationarity test on the collected data. It includes two methods. One is to check the stationarity by looking at the graph. The other is to check the stationarity by unit root. For the unstable data, we make the time series smooth by difference. The first and second-order difference formulas are as follows:

$$f'(x_k) = \frac{f(x_{k+1})+f(x_{k-1})}{x_{k+1}-x_{k-1}} \tag{5}$$

$$f''(x_k) = \frac{f(x_{k+1})+2f(x_k)+f(x_{k-1})}{h^2} \tag{6}$$

After the stationarity test, we will identify the ARIMA model. We need to identify the model by judging the order $p$ of the autoregressive model and the order $q$ of the moving average model. According to the values of p and q, the ARMA model is defined as follows:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{7}$$

Where $y_t$ is the current value. $\mu$ the constant term. $p$ and $q$ are ordered. $\gamma_i$ is autocorrelation coefficient. $\theta_i$ is the moving average coefficient. $\epsilon_t$ the error.

During the process of parameter estimation, we adopted Bayesian Information Criterion (BIC) criterion to solve the problem of choosing $p$ and $q$ parameters in ARIMA, so that the fitting effect of ARIMA can be optimized. BIC makes up for the deficiency that AIC does not converge to the actual model when the sample size is large. The definition of BIC with n samples is:

$$BIC = kln(n) - 2ln(L) \tag{8}$$

**(2) The Process of Model**

To make the data fitting better, we used the ARIMA model to forecast the global average temperature. We got the average monthly temperature for all regions of the world from 1743 to 2013 from Kaggle. First, we processed the null values and outliers in the data set and then used the complete dataset to calculate the global annual mean temperature from 1743 to 2013 and draw the global mean temperature time series diagram.

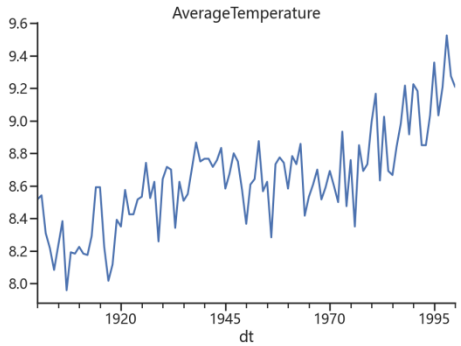



**Figure 6.** The Sequence Chart of Global Average Temperature.

From Figure 6, we can see that from 1943 to 2013, the global mean temperature showed a fluctuating rising trend. It means that global temperature data based on time series are unstable and cannot be directly predicted by models. Therefore, we use the 1-order and 2-order difference methods to stabilize the global mean temperature data, and the processing results are shown in the figure below.

As shown in Figure 7, the blue line shows the original graph without a difference, while the orange and green lines show the results of a 1-order difference and a 2-order difference. Subjectively, the orange and green lines fluctuate around the value of 0. After carrying out the unit root test, we found the result was less than 0.01. Therefore, the results of the first-order difference and the second-order difference were both stationary sequences. In fact, time series are already close to stationary series at 1-order difference. After that, we calculated the autocorrelation coefficient and partial correlation coefficient of the model and drew the average temperature autocorrelation chart and partial correlation chart.

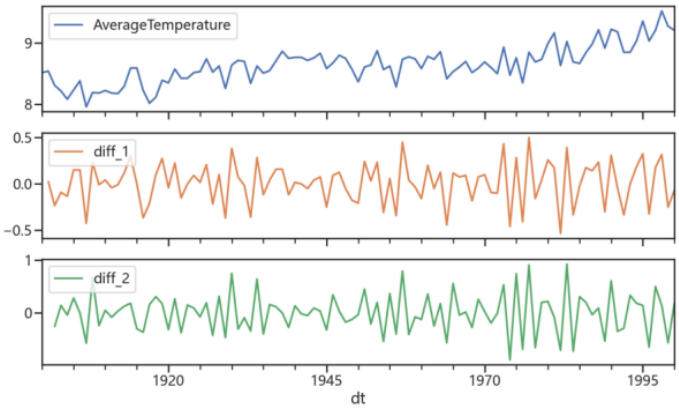
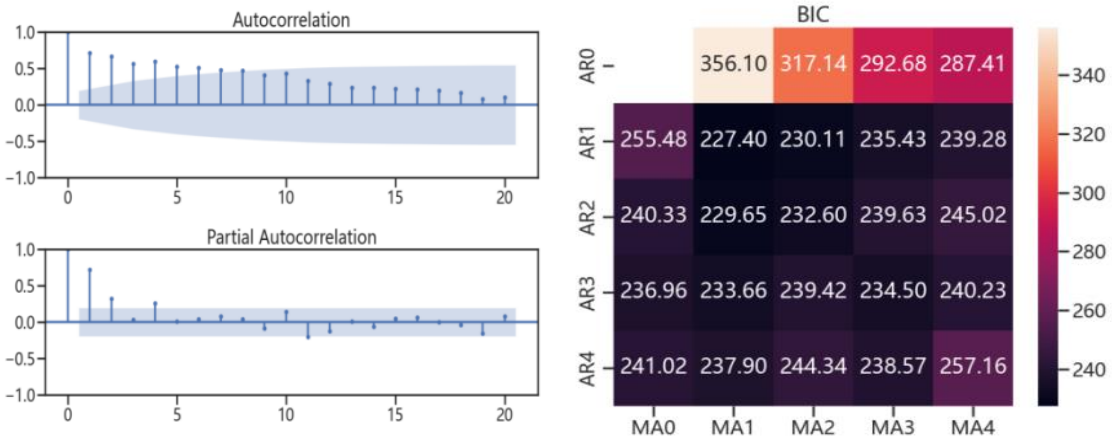



**Figure 7.** The Result of the Difference.

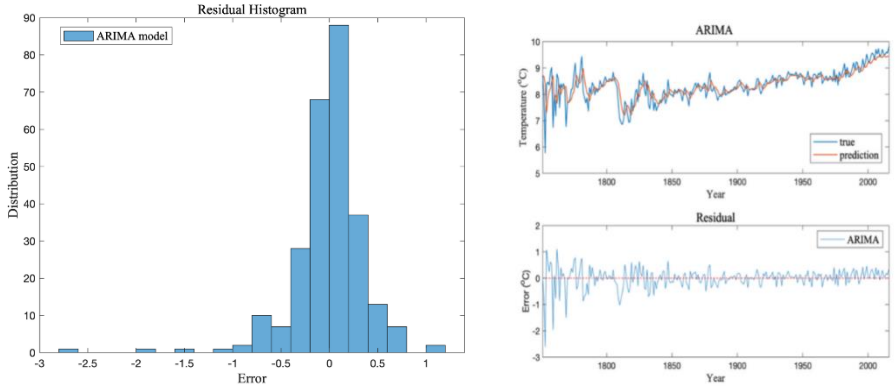



(a) (b)

**Figure 8.** Autocorrelation and Partial Correlation and The Result of BIC.

The values of $q$ and $p$ cannot be accurately determined due to the subjectivity of identifying the model through the senses. Therefore, the BIC criterion is adopted to judge the optimal model and determine the values of $p$ and $q$. It can be seen from Figure 8(b) that when the values of $p$ and $q$ are both 1, the value of BIC is the largest, and the model fitting effect is the best.

### (3) Model Test and Result




(a) (b)

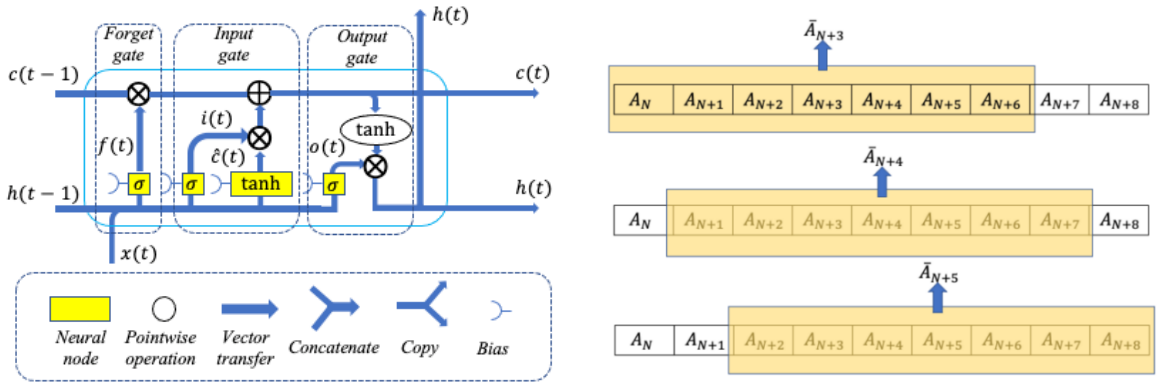
**Figure 9.** Residual Histogram of ARIMA and The Results of ARIMA Model Fitting.

The standardized residual histogram shows that the left and right sides are symmetrical, and the scatter is basically close to a straight line, as seen in the P-P plot. It indicates that the residual normality results are promising. Moreover, the results of the model fitting are in Figure 9(b). It is obvious that the temperature of predicted is in line with the true data. ARIMA model fits well.

### 2.2.3 Temperature Prediction Model based on LSTM

#### (1) Description of LSTM Model

LSTM [3] consists of an input, hidden, and output layer. The hidden layer is composed of one or more memory units. Each memory unit has three gates: forgot gate, the input gate, and the output gate.




(a) (b)

**Figure 10.** The LSTM Structure Chart.

As shown in Figure 10(a), the line at the top of the memory unit represents the propagation path of cell state $c$, which traverses the entire unit. Firstly, it relates to the input, forgetting, and output gate through some linear computation operations. Secondly, it combines the memory of LSTM at the current time $t$ with the previous time $t-1$ to form a new cell state $c(t)$. The cell state $c(t)$ can realize timely, effective, and unchanged information propagation in the memory unit by using a small amount of linear interaction. LSTM involves the following formula:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f \tag{9}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i \tag{10}$$

$$\bar{c}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{11}$$

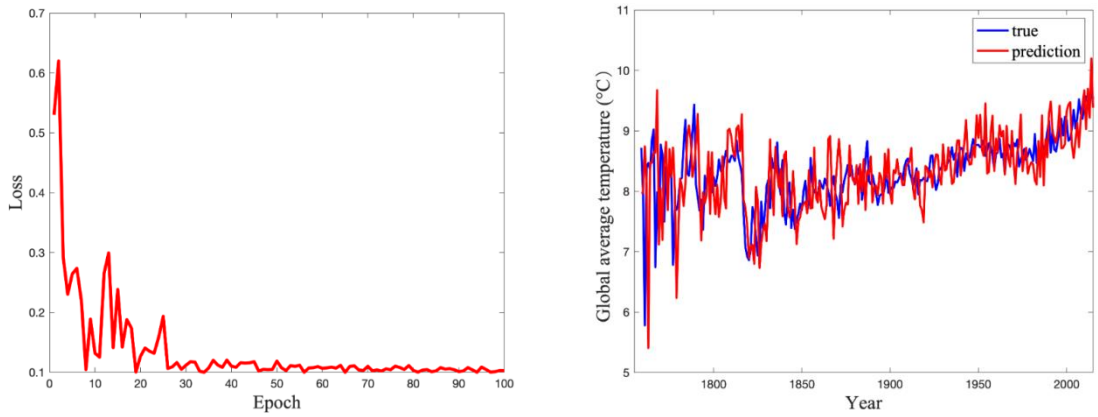
$$c_t = f_t \otimes c_{t-1} + i_t \otimes \bar{c}_t \tag{12}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{13}$$

$$h_t = o_t \otimes tanh(c_t) \tag{14}$$

Where $f_t$ is, the forgot gate at time $t$. $i_t$ is the input gate at time $t$. $c_t$ is the output gate at time $t$. $\bar{c}_t$ is the cell state of the current input.$W_f$ and $b_f$ represent the weight matrix of the forgot gate and bias. $W_i$ and $b_i$, respectively, represent the input gate's weight matrix and bias. $W_c$ and $b_c$ represent the weight matrix of cell state and bias. $[h_{t-1}, x_t]$ matrix comprises the output at time $t-1$ and the input at time $t$. $\sigma$ and $tanh$ are the sigmoid functions and tanh function. Symbol $\otimes$ means that vectors multiply by elements.

Shown in Figure. 11(a), the model loss rate decreases as the number of iterations increases. After more than 30 iterations, the model loss rate is almost 0.1. The error of our model is small, and the model effect is good. Figure 11(b) compares the predicted results of LSTM model and the actual values. The blue line represents the accurate global average temperature, and the red represents the LSTM forecast. The trends of the two lines are almost identical. The two lines are very consistent, both in terms of the wave shape and the uptrend. It shows that our model has an excellent fitting effect, providing sufficient data support for our future temperature predictions in 2050 and 2100.




(a) (b)

**Figure 11.** The Loss and the Predicted Results of LSTM.

### 2.3. The Solution of Model

**Table 1.** The Predict Result

| model \ year | 2050 | 2100 | | The year reached 20 ℃ | |
|---|---|---|---|---|---|
| Ridge Regression | 10.2775℃ | 11.5168℃ | | 2398 | |
| ARIMA | 9.476℃ | 9.211℃ | | / | |
| LSTM | 10.231℃ | 11.591℃ | | 2439 | |

We use three time series, as mentioned earlier, forecasting models we established to predict temperatures for two years, 2050 and 2100, with the outcomes shown in the following table 1. At the same time, we answer in which year the Earth's average temperature will exceed 20°C. The results are shown in the flowing table.

It should be noted that the ARIMA model only supports short-term predictions, as it lacks a future and, therefore, cannot predict a specific value. However, under the assumption of a smooth and slow decline in temperature, the ARIMA-based model predicts that future temperatures will not reach 20°C.

$$T_{pred} = 1.3725 \cdot 10^{-8}x^3 - 2.7633 \cdot 10^{-5}x^2 - 0.0378x + 85.7242 \quad (15)$$

### 2.4. Evaluation of Time Prediction Model

In conclusion, the LSTM model is the most reliable in terms of accuracy. Still, the need for subsequent training of updates to the dataset makes predicting global temperatures up to 20 degrees Celsius less effective, like the ARIMA model. The predicted outcomes of the polynomial fit are the worst as it is affected by the extreme data and performance[4-6].

Although the regularized polynomial regression model avoids the problem of underfitting and overfitting the fitted curve, it is still sensitive to data noise and is greatly influenced by extreme data. For example, the irregular variation of temperature before 1800 makes the fitting effect not optimal, which makes the loss value of the appropriate process more extensive; between about 1800 and 1900, the temperature has a slightly increasing trend but increases slowly, which makes the curvature of the fitted curve significantly smaller; the rapid increase after 1970 leads to a fast-increasing trend of the fitted curve again[7]. From the results, the global average temperature will reach 20 degrees Celsius in just over two hundred years, indicating that the model is affected by the dramatic growth of the data set in the latter half of the period, and the fit is average. Time series analysis method more accurately reflects their trends, and the computational complexity is low (the calculation is not complicated), but the accuracy is limited. Furthermore, it needs subsequent data updates. Otherwise, long-term prediction cannot be achieved. The dataset must be stable if we want to use the ARIMA model to predict the time series data. Otherwise, it is impossible to capture the pattern. However, the

temperature is non-stable (fluctuates wildly) in the long term due to unexpected factors such as forest fires, volcanic eruptions, epidemics, and industrial revolutions. Even if we use a different method to stabilize the temperature, it will lead to inaccurate prediction also. In addition, although we used the sliding window method to avoid the long-term prediction convergence to the mean in this model, the predicted values still show a steady trend of flat decline. Machine learning models, such as LSTM networks, have high computational complexity and high accuracy but can only make short-term predictions because they are also data-driven methods. A large amount of data is required to train the model, and subsequent data updates are needed to fill it. Otherwise, long-term prediction cannot be achieved[8-10].

## 3. Conclusions

In order to assess past temperature changes, this paper first constructs a growth rate indicator, which is the ratio of data growth over a certain period of time. Firstly, since the growth rate before 1800 is not significant, this study only analyses the growth rate changes after 1800. In addition, this study uses the Mann-Kendall method to determine the trend of the growth rate. Next, ridge regression model, ARIMA model and LSTM model were constructed to predict the future temperature changes. The results show that the LSTM model is the most reliable, while the ARIMA model is only suitable for short-term prediction due to the lack of follow-up datasets and is less effective in predicting global temperatures up to 20°C. The results of the ARIMA model are summarised below. The polynomial fitting results were affected by extreme data and performed the worst. The results indicate that the temperature will reach 20°C around 2310 and within 2398~2439 years based on the ridge regression model and LSTM model, respectively.

## References

[1] Cao J P, Chi DC, Wu L Q, Liu L, Li S Y and Yu S. "Manner-Kendall Examination and Application in the Analysis of Precipitation Trend.", Agricultural Science & Technology and Equipment 5(2008): 35–40.

[2] Wang Y H. "Global Temperature based on ARIMA Model and LSTM Neural Network Analysis of Prediction." Science and technology Innovation 35(2021): 166-170.

[3] Tao Ye, Du Jinglin. "Temperature Prediction Using Long Short Term Memory Network Based on Random Forest." Computer Engineering and Design 3(2019): 737–743.

[4] Li G Y, Fu Q, Sun Y and Feng Y. "Grey Correlation Analysis Model based on Entropy Weight and its Application." Journal of Water Resources and Water Engineering 17(2006): 4.

[5] Zhu H Y, Zhang S X, Liu J Z, Zhang Y R and Wang T. "Spatial-temporal Characteristics and Influencing Factors of Air Quality in Shandong Province." Environmental Pollution & Control 44(2022): 1228-1234.

[6] Luo Y, Li Y L. "Comprehensive Decision-Making of Transmission Network Planning Based on Entropy Weight and Grey Relational Analysis." Power System Technology 37(2013): 77-81.

[7] ZHUGE J W, Wang D W, Chen Y, Ye Z Y and Zou W. "A Network Anomaly Detector Based on the D-S Evidence Theory." Journal of Software 17(2006): 463-471.

[8] M. V, A. N. Resilience and vulnerability to climate change in the Greek Dark Ages [J]. Journal of Anthropological Archaeology,2021,61.

[9] Kai W,Junguo L,Jun X, et al. Understanding the impacts of climate change and socio-economic development through food-energy-water nexus: A case study of mekong river delta [J]. Resources, Conservation & Recycling,2021,167.

[10] Ali A C,Yuansheng J,Waqar A, et al. Addressing the effect of climate change in the framework of financial and technological development on cereal production in Pakistan [J]. Journal of Cleaner Production,2021,288(prepublish).