Study of Global Temperature Levels Based on the ARIMA Model

Yu Yuan [1,*,†], Haidong Huang [2,†], Shuangquan Wen [3,†]

[1] School of Statistics, Chengdu University of Information Technology, Yibin, China, 610103

[2] College of Communication Engineering, Chengdu University of Information Technology, Chengdu, China, 610225

[3] School of Electronic Engineering, Chengdu University of Information Technology, Chengdu, China, 610225

* Corresponding Author Email: 13458473946@163.com

† These authors contributed equally.

Abstract. The continuous growth of global carbon dioxide content has exacerbated global warming [1]. Based on this issue, this article mainly studies global temperature levels based on the ARIMA model. This article first uses data collection and preprocessing to determine whether the global temperature increase in March 2022 has led to a greater increase than the past 10 years. Then, ARIMA and seasonal ARIMA are established to describe the past and predict the future, predict the specific time when the average temperature reaches 20℃, and test the accuracy of the model. Finally, it is concluded that seasonal time series prediction is more accurate. For this study, this article first collects and completes the required temperature data for visual analysis. And independent sample t-tests were conducted on the two rising scenarios, and the results did not lead to a greater increase. Then the chapter describes the past data through the Run chart, and uses ARIMA and seasonal ARIMA to predict the future global temperature level [2]. Afterwards, this article adopts pre analysis on the data used and uses a predictive model to calculate the specific time to reach 20℃. The study found that the average temperature could not reach 20.00℃ in 2050 or 2100, but reached 20℃ in 3062. Finally, based on the data description and the characteristics of the data itself, this article demonstrates that seasonal ARIMA is more accurate.

Keywords: Data Collection and Preprocessing, Independent-sample T-test, Seasonal Time Series.

1. Introduction

Global warming is caused by the imbalance of energy absorbed and released by the Earth's atmospheric system, as well as the continuous accumulation of energy in the Earth's atmospheric system.

Human activities, such as the burning of fossil fuels and deforestation, produce large amounts of greenhouse gases such as carbon dioxide. Visible light from the solar radiation reaches the surface of the earth through the atmospheric greenhouse gases and is absorbed by the surface. When the surface is cooled, the light absorbed by this part is radiated into the air by the earth in the form of infrared radiation, and the infrared ray with long wave energy is absorbed by the greenhouse gases in the atmosphere. In this way, the greenhouse gases in the earth's atmosphere are like the glass of the greenhouse. Sunlight can shoot into the greenhouse, but the heat in the greenhouse cannot be emitted, resulting in the "greenhouse effect" [3]. The "greenhouse effect" caused by increased concentrations of atmospheric greenhouse gases is the cause of global warming. Global warming will trigger a series of natural disasters, including rising sea levels, the spread of infectious diseases, the increasing number of hurricanes, and the increasingly frequent floods and droughts, which will cause great damage to human survival [4].

This paper is mainly to build a differential and diagnostic model of global warming, which specifically needs to study the following aspects:

First through data collection and preprocessing, determine whether the rise in global temperature in March 2022 lead to greater rise than during the past 10 years, then establish a time series [5] and

seasonal time series model [6] to describe the past, and predict the future global temperature, on the basis of the average temperature of 20℃ specific time, and test the accuracy of the model. Finally, the study conclusions were drawn.

2. Establishment and solution of the global temperature level research model

2.1. Establishment of the global temperature level research model

2.1.1 Model hypothesis

1) Suppose that the relative humidity is constant and the particle evaporation speed is similar in each region;

2) Suppose that the weather sampling data in each region is true and effective;

3) It is assumed that the acquired data has a certain correlation and has an obvious periodicity;

4) Suppose that other factors such as terrain and ground conditions on the climate are not considered.

2.1.2 Solution of the ANOVA analysis

(1) Data collection and preprocessing

Then, the study needs to preprocess the data.

First, since the existing data ended to 2013, it is impossible to determine whether the rise in global temperature in March 2022 led to a larger increase than in the past 10 years period. Data should be collected at http://berkeleyearth.org/data/ to obtain monthly data of global temperature and aggregated annual data tables.

By illustration of this data, the study concluded that the acquired data is actually based on a relative temperature of 1951.1-1980.12 temperature. This paper intends to use monthly temperature anomaly to directly describe the temperature trend. Therefore, the data table can be obtained by directly importing the aforementioned monthly temperature anomaly form into excel, and by deleting and other processing;

At this point, the data is completed.

(2) Visual analysis of the —— preliminary grasp

In order to study whether the pairing of global temperature in March 2022 led to a greater rise than in the past 10 years, this paper first uses visualization to grasp the overall trend as follows, and as shown in Figure 1:

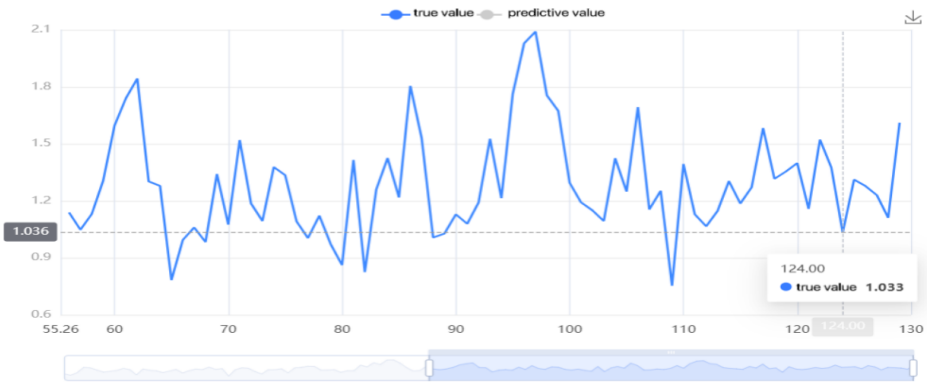


Figure 1. March 2022 versus the last 10 years

It can be seen intuitively that after March, there is no significant upward trend, or even a downward trend. However, this is a preliminary analysis, and specific differential analysis is needed later.

(3) Difference analysis —— actual calculation

Independent sample T-test [7]: Independent sample t-test (no correlation between each experimental treatment group, namely, independent sample), which is used to test the difference in the data obtained by the two groups of unrelated samples.

The independent sample t-test statistics is:

$$t = \frac{\overline{X1} - \overline{X2}}{\sqrt{\frac{(n1-1)S1^2+(n2-1)S2^2}{n1+n2-2}\left(\frac{1}{n1}+\frac{1}{n2}\right)}} \quad (1)$$

$S1^2$ and $S2^2$ are the two-sample variances; n_1 and n_2 are the two sample sizes.

Analytical procedure

1) Group the quantitative fields (Y) according to the classification variables (X), and test the normality test respectively to check whether the overall distribution of the data shows a normal distribution. If the test passes, it is recommended to use the independent sample T-test.

2) Check the Mann Whitney test table, if it is significant, you can see the median to analyze the difference, otherwise it indicates no difference.

3) If the Mann Whitney-test of independent samples is significant, the difference can also be quantified by effect quantitative analysis.

Detailed conclusions

Output 1: normality test results

Table 1. Normality test results

Variable name	Sample size	Average	Standard deviation	Skewness	Kurtosis	S-W test	K-S test
Monthly Anomaly	130	1.202	0.34	0.372	0.86	0.977 (0.028**)	0.075 (0.443)

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

Chart description: As shown in Table 1, quantitative variable Monthly Anomaly data conform to normality. And the overall distribution of the data presents a normal distribution, so the independent sample T-test can be used.

Output 2:

Table 2. Table of results of Mann Whitney U test analysis

Variable name	Variable value	Sample size	Median	Standard deviation	Statistical quantities	¥	Median value difference	Cohen's d value
Monthly Anomaly	1	123	1.158	0.346	337	0.335	0.12	0.237
	2	7	1.278	0.188				
	Total	130	1.169	0.34				

Note: ***, ** and * represent 1%, 5% and 10% significance levels respectively

Chart description: As shown in Table 2, the above table shows the results of the independent sample Mann Whitney U-test, including the median, statistic, and effect size Cohen's d values. P-values for each analysis analyzed were not significant.

At this point, it can be concluded that the rise in global temperature in March 2022 is not significantly different from the increase in the past 10 years, consistent with the visual intuition.

2.1.3 The determination of the number of network layers

The collected data was processed as in the previous article to obtain the data sheet.

(1) Describe the past

Here, you can visually see the visualization of the historical data. You can see that the global temperatures are constantly rising. As shown in Figure 2.

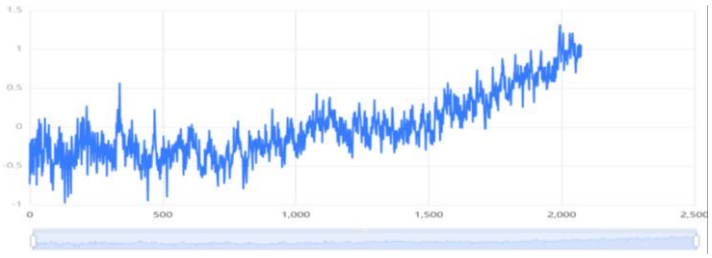


Figure 2. Visualization of global temperature history data before 2022

(2) Predicting the future —— time-series models

We use two models for future forecasting, the first of which is the ARIMA time series model.

ARIMA time series model: It is a model built by transforming a non-stationary time series into a stationary time series and then regressing the dependent variable on its lagged values only and on the present and lagged values of the random error term.

Analysis steps

1) The ARIMA model requires the series to satisfy smoothness. View the results of the ADF test and analyse whether it can significantly reject the hypothesis that the series is not smooth ($p<0.05$) based on the analyzed t-values.

2) View the comparison graph of the data before and after differencing to determine whether it is smooth (not much up and down fluctuation), and also bias the time series (autocorrelation analysis) and estimate its p and q values according to the truncated tails.

3) The ARIMA model requires the model to have pure randomness, i.e. the model residuals are white noise. Check the model test table and test the model white noise according to the p-value of the Q statistic ($p>0.05$), and derive the model formula combined with the time series analysis graph for a comprehensive analysis to obtain the backward prediction order results [8].

Detailed findings

Output 1: ADF test table

Table 3. ADF test table

Variables	Differential orders	t	¥	AIC	K		
					1%	5%	10%
Monthly Anomaly_Outlier Handling	0	-1.028	0.743	361.014	-3.438	-2.865	-2.569
	1	-10.384	0.000***	355.8	-3.438	-2.865	-2.569
	2	-12.139	0.000***	457.388	-3.438	-2.865	-2.569

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

Chart description: As shown in Table 3, the results of the ADF test indicate that the sequence is smooth time series data. The null hypothesis of sequence instability can be significantly rejected by analyzing the t-values.

Output 2: Time series plot

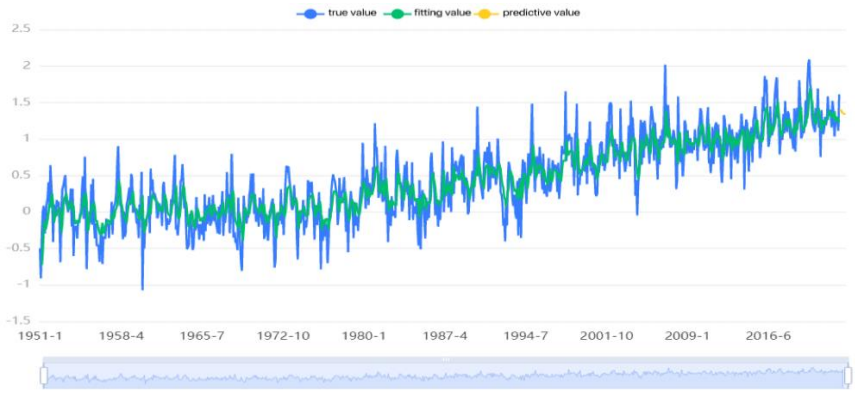


Figure 3. Time series chart

Chart description: As shown in Figure 3, the graph above represents the raw data plot, model fitted values, and model predicted values for this time series model.

(3) Forecasting the future - seasonal time series models

Seasonal ARIMA forecasting model: A time series with cyclical variations caused by seasonal or other cyclical factors is what we call a seasonal time series and the corresponding model is a seasonal model. Some time series data will exhibit a certain degree of cyclicality [9].

Analysis steps

1) Decompose the time series into trend data in order to make a preliminary determination of the seasonal effects of the data.

2) The seasonal ARIMA model requires the time series to satisfy the smoothness test, if P<0.05, the series is smooth. If the original time series does not satisfy the smoothness, it is differenced as well as seasonally differenced until the series satisfies the smoothness.

3) view the final difference series plot, while biasing the time series (autocorrelation analysis) and estimating its p and q values based on the truncated tails.

4) The ARIMA model requires the model to be purely stochastic, i.e. the model residuals are white noise. Check the model test table, based on the P-value of the Q-statistic (a P-value greater than 0.05 is white noise); check the degree of fit of the model to the series R^2, the closer to 1 the better the model effect.

Detailed conclusions

Output 1: sequence decomposition Figure [10]



Figure 4. Sequence decomposition map

Chart description: As shown in Figure 4, the overall direction of long time series data is increasing, with a seasonal effect.

Output 2: Time series plot

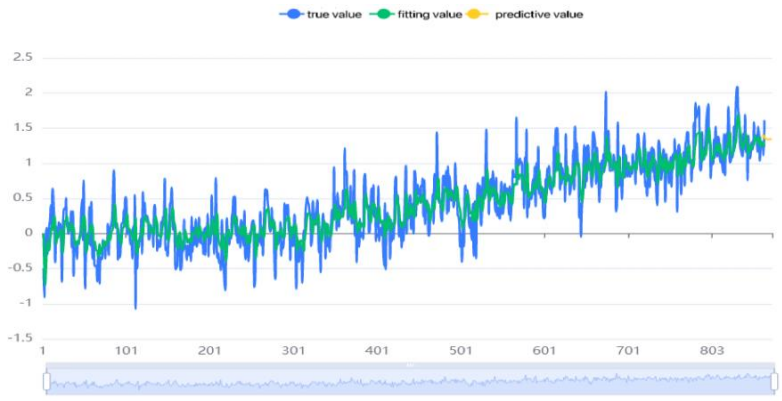


Figure 5. Time series plot

Chart description: As shown in Figure 5, the above figure represents the original data map, model fit value, and model prediction value of this time-series model.

2.2. Solution of the global temperature level research model

(1) Antecedent analysis

Prior analysis is required before actual forecasting can take place.

The time series forecasting model does not predict well due to too much data. So this question should change the monthly data to annual data to make predictions for the 2050 and 2100 data. And we based the absolute temperature on the estimated 1951.1-1980.12: 8.60 +/- 0.05, so the pre-1951 data needs to be removed. The large data prediction model for the user's electricity consumption is implemented in the Clementine software.

(2) Time series forecasting

Since the feasibility of both models has been demonstrated, the prediction results are directly presented here.

1) 2050 forecast: the predicted value is 1.36, and with our relative temperature, the real temperature can be 10.176; less than 20℃.

2) 2100 forecast: the predicted value is 1.45, plus the relative temperature, still less than 20℃.

(3) Seasonal timing forecast

1) Seasonal time forecast for 2050: the predicted value is 1.37, plus the relative temperature is less than 20℃.

2) Seasonal time forecast in 2100: Since the maximum forecast unit of seasonal time series forecast is 48, it is forecast twice, and the final forecast value is 2.110, plus the relative temperature is less than 20℃.

(4) Time of up to 20℃

After many predictions, the seasonal time forecast can not meet the seasonality, so the time series is used to predict at this time.

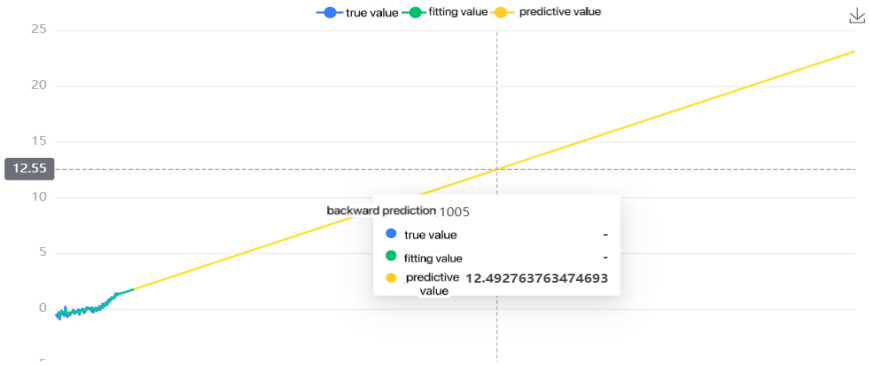


Figure 6. Time to reach 20°C

As shown in Figure 6, 1005 units need to be predicted backwards, that is, 3027 to reach the average temperature of 20℃.

3. Conclusion

1) The global temperature increase in March 2022 compared with the increase in the past 10 years, the independent sample t-test was conducted, and through the specific calculation test, the final result did not lead to a larger increase.

2) In 2050 or 2100, the average temperature was not 20.00℃ in the prediction model, but 20℃ in 3062.

3) For the two models used here, the study shows that seasonal time-series predictions are more accurate. The reasons are as follows: The data in this study have clearly stated that since the earth's temperature is not symmetrically distributed around the equator, the mean global land temperature intake is average seasonal. The website report presented in the present study also indicates the existence of seasonal variation.

References

[1] The US National Oceanic and Atmospheric Administration. The National Oceanic and Atmospheric Administration Earth Systems Research Laboratory. (October, 2022). Trend of carbon dioxide in the atmosphere.

[2] Wang Yuanhao. Analysis of global air temperature prediction based on the ARIMA model and the LSTM neural network [J]. Science and Technology Innovation, 2021 (35): 166-170.

[3] Xu Sichuan, Ma Xiyu, Zhou Tianhao, etc. Carbon emission CO_2 greenhouse effect mechanism [J]. Journal of Yunnan University (Natural Science Edition), 2023, 45 (02): 513-522.

[4] The US National Oceanic and Atmospheric Administration. Research news and specialty articles in NOAA. (June 3, 2022). Carbon dioxide is now more than 50% higher than pre-industrial levels.

[5] Snow White. Temperature prediction and heat heat wave warning analysis based on time series model and XGBoost combination [D]. Nanjing University of Information Science and Technology, 2021.

[6] Dong Yanyan. Research on the construction of ARIMA-BiLSTM model in vegetable price prediction under different combination modes [D]. Dongbei University of Finance and Economics, 2022.

[7] Yang Weikang, Qiao Jianfang, Gao Xingyi, Zhong Wenqin. Use of the rank-sum test in the analysis of bird habitat selection [J]. Drought Zone Study, 2001 (03): 42-44.

[8] Wang Yan. Applied a time-series analysis [M]. Beijing: China Renmin University Press, 2005.

[9] Hao Junzhang, Cui Yujie, Han Jiangxue. Passenger volume prediction based on ARIMA model in China [J]. The Practice and Understanding of Mathematics, 2015 (18): 10.

[10] Wang Fang, Yu Lean, Zha Rui. Integrated prediction modeling of scale decomposition of electronic waste recovery driven by seasonal data characteristics [J]. Chinese Management Science, 2022, 30(03): 199-210.