Proceedings of the 4th International Conference on Materials Chemistry and Environmental Engineering
DOI: 10.54254/2755-2721/58/20240725

Factor analysis and countermeasures for global warming based on data science methods

Yuning Man

School of Science, Hangzhou Dianzi University, Zhejiang, Hangzhou, China,310018

1714019559@qq.com

Abstract. In recent years, the frequent and intensifying occurrence of extreme weather events such as heatwaves, floods, and storms worldwide has led scientists and policymakers to be increasingly concerned about the phenomenon of global warming. The global average temperature has steadily risen over the past few decades, and 2015-2022 has been identified as the warmest years on record. Taking global warming as the research object, the main problem is to find out the factors affecting climate change. The article predicts the global average temperature of the future and builds a model to analyse the relationship between global average temperature, time, location and the factors affecting climate change. The prediction models of ARMIA and EEMD-LSTM were built, and the model with the best fit was obtained after several iterations of tuning and trial optimization. The results showed that Flood, Storm, and Extreme temperature, had a high correlation with the global average temperature. The data of greenhouse gas indicators such as CO2, NO, and CH4, and catastrophes such as earthquakes, Volcanic activity, and wildfires have objective effects on global warming. Random forest regression model pairs were developed, and analysis of the importance of each component to the model showed that CO2 concentration and CH4 concentration had significant effects on global average temperature. Propose measures that can curb or slow down global warming.

Keywords: Climate Warming, LSTM, ARIMA, Correlation Analyse.

1. Introduction

Excessive emissions of greenhouse gases lead to the accumulation of the greenhouse effect, which leads to an increase in the temperature of the earth and leads to global warming. Since 2023, the world has experienced unusually hot weather and extreme temperatures, several countries in the northern hemisphere have experienced persistent heat and heat wave weather, and many parts of Europe have broken record records for high temperatures. At the same time, global warming could lead to more extreme phenomena such as rising sea levels, melting glaciers, and more, which has caused widespread concern for the issue of global warming. To further understand global warming trends, models were built based on historical data on global temperature, aiming to contribute more to the process of slowing global warming.

According to the Restatement of the Problem, the work mainly includes the following:

Based on global climate change data, ARIMA [1] and LSTM [2] models are established to describe the past and predict the future of global temperature levels;

DOI: 10.54254/2755-2721/58/20240725

The influence of natural disasters, time, and region on global temperature levels and their interrelationships are explained, considering the impact of natural disasters on temperature, and measures to cope with global warming are proposed;

A non-technical report was written to describe the models developed and to suggest possible options and recommendations for mitigating global warming. The global average temperature is predicted to reach 14.15 ℃ in 2050 and 14.51 ℃ in 2100 by ARIMA, reach 14.30 ℃ in 2050 and 15.60 ℃ in 2100 by EEMD-LSTM; LSTM predicts that the global average temperature will reach about 20 ℃ in 2200, while ARIMA will reach 2860.

Pearson correlation analysis was performed, combined with the latitude and longitude of cities, and the results responded that cities near the equator are more associated with global warming, while city-states at higher latitudes are less associated with global warming.

12 representative disaster indicators were selected in the past 50 years and analyzed the grey correlation with the global temperature data.

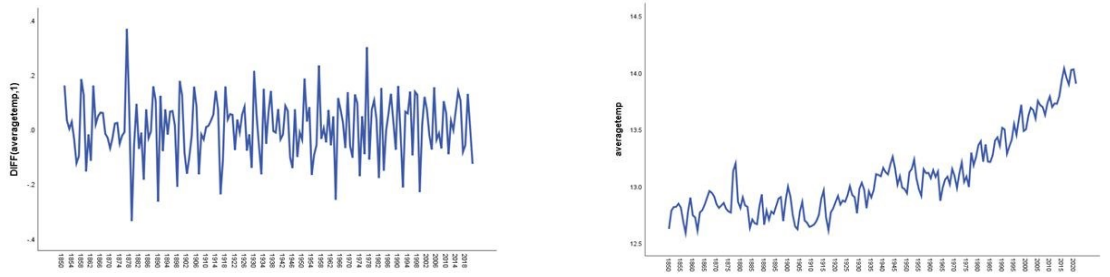
In order to avoid the influence of the uncertainty of the observations on the predictions, the GMST data from Berkeley Earth are used to perform the fitting tests, modelling and prediction tests. ERSST V5 data were also used for the calculation of three climate indices (Ni$\tilde{\text{n}}$o3.4, IPO, AMO) and the attribution analysis of GMST.

2. Model 1: ARIMA

ARIMA's full name is the Autoregressive Integrated Moving Average Model. It is also demonstrated as follows:

$$y_t = \sigma + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots\cdots + \theta_p y_{t-p} + \phi_1 q_{t-1} + \phi_2 q_{t-2} + \cdots\cdots + \phi_p q_{t-p} \quad (1)$$

The ARIMA model was applied to the year-by-year global temperature to make a series plot and a primary difference plot as shown in Figure 1.



(a) Global average temperature series(left)

(b) Global temperature primary differential results(right)

Figure 1. Data Screening.

There is a significant improvement in the smoothness of the series plot after one difference, which confirms that the global land-ocean mean temperature has a gradually increasing trend.

By using SPSS, smoothness, periodicity, autocorrelation and bias correlation analyses were performed for year-by-year data to provide guidance for the construction and parameter optimization of the ARIMA model. After the analysis and several parameters

tuning, the model fits and predicts best when the ARIMA (0, 1, 2) model is used for fitting and forecasting the month-by-month data. The R-squared of the model is 0.884, and the model results have good confidence.

The fitting and prediction results are given in Figure 2

DOI: 10.54254/2755-2721/58/20240725

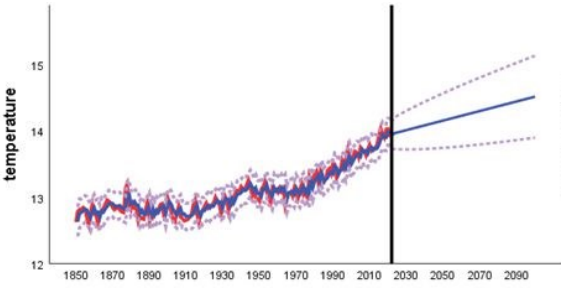


Figure 2. ARIMA Forecast Results.

3. Model 2: EEMD-LSTM

3.1. EEMD Decomposition

EEMD [3] is a noise-assisted data analysis method that addresses the shortcomings of the EMD method. The principle of EEMD decomposition is that when the additional white noise is uniformly distributed throughout the time-frequency space, the time-frequency space consists of different scale components split by filter banks. When the signal is coupled with a uniformly distributed white noise background, the signal regions of different scales are automatically mapped to the appropriate scale associated with the background white noise. Of course, each independent test can produce very noisy results, due to the fact that each additional noise component includes both the signal and the additional white noise. Since the noise is different in each individual test, the noise will be eliminated when the full mean of the tests is used. The full mean will eventually be considered the true result, and as more and more tests are performed, the additional noise is eliminated and the only lasting solid component is the signal itself.

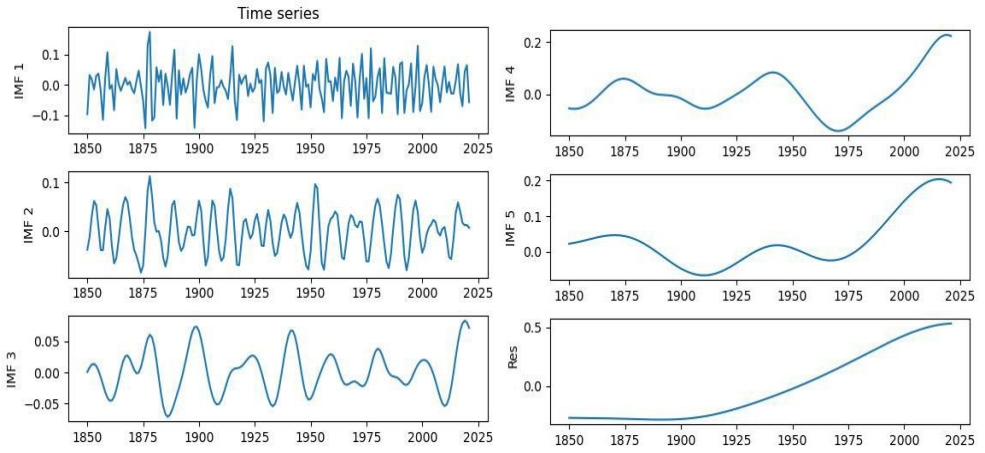


Figure 3. EEMD decomposition results.

Taking Berkeley Earth's data as an example, the average GMST from 1860 to 2019 was selected for EEMD decomposition, and the best decomposition result was determined after pre-experiments in order to eliminate the noise of the original series with an additional noise amplitude of 0. 1 and set number of 45 when the decomposition result has the smallest root mean square error with the original series. Figure 3 shows the five eigenmodal components (IMF1-5) and the residual term (Residual), which can be considered as the long-range trend of the whole sequence. EEMD decomposes

the original sequence into subseries of different scales, from high to low frequencies, and all with quasi-periodic variations. The final residuals are also consistent with the long-range trend of global warming. The final residuals are also consistent with the long-range trend of global warming. For the above decomposition of the IMF, rescaled range analysis, also known as R/S analysis, is used to

DOI: 10.54254/2755-2721/58/20240725

calculate the Hurst index of each subseries and is a good time series indicator. When 0.5¡Hurst¡1, the Hurst index is a good indicator of the long-range correlation of the time series When 0.5¡Hurst¡1, the series has long-term memory, i.e., the closer to 1, the stronger the front-to-back correlation of the series. When 0.5¡Hurst¡1, the sequence has long-range memory, i.e., the closer to 1,

the stronger the pre-post correlation of the sequence, and vice versa, the closer to 0.5, the worse When Hurst=0.5, the sequence is a completely random sequence. the Hurst indices of IMF2 to 6 are all much higher than 0.5, and only IMF1 is 0.6238. is 0.6238, indicating that the IMF1 long-range correlation is much weaker than the long-range correlation of the other subsequences.

3.2. Modeling EEMD-LSTM

Based on the EEMD decomposition, the LSTM approach is used to build the prediction model. The LSTM model is derived from the traditional RNN [4] optimization, which changes the shortcomings of the RNN circular gradient and low utilization of historical data, and is itself a multilayer neural network, which also has better prediction ability for nonlinear sequences. its network structure is shown in Figure 4.

In establishing the LSTM model, for IMFs [5] with different change cycles, the corresponding neural networks should be established to predict them separately. different

subsequences of the LSTM model refer to the before-and-after correlation of their own sequences, with the training step and the number of features as the input dimension, the number of input nodes as the training set, the number of output nodes as the prediction set, and the number of hidden nodes as the number of hidden layers. In the same IMF model, the above parameters are generally fixed values, while the learning rate, number of iterations, gradient descent rate, etc. are adjusted according to the complexity of each IMF to ensure accurate prediction with the shortest possible training time, due to the existence of random weights and random initialization of LSTM.

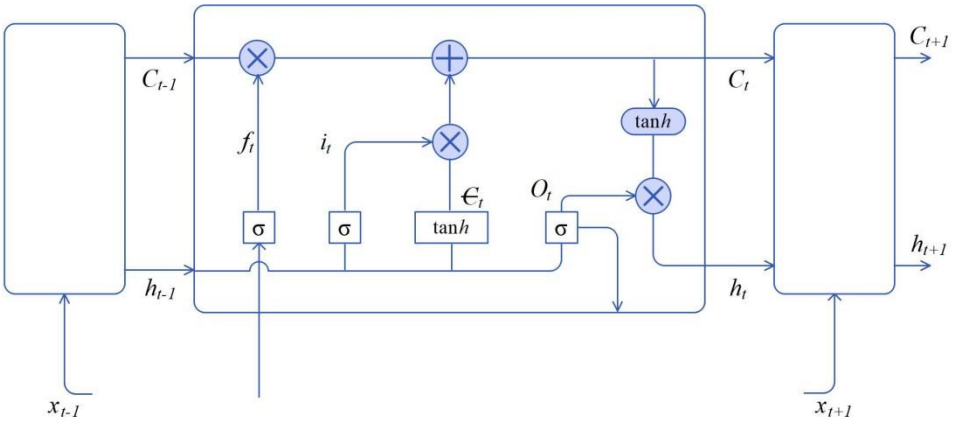


Figure 4. Single-layer LSTM neural network structure.

After 20 replicate test means, the memory lengths of IMFs 1 to 5 were identified as 10, 25, 25, 10, and 7 years, respectively, with the goal of minimizing the goodness-of-fit, and significance tests were performed. The final individual MFs fit and prediction plots are shown in Figure 5. As shown in Figure 8 both perform almost identically on the last four IMFs, indicating that future changes at the time scales represented by IMF3-6 are mainly influenced by historical changes, and such changes can be predicted more accurately after sliding difference, while it is difficult to fit IMF1 better with either model due to the low long-range correlation of IMF1, and the Pacific and Atlantic climate oscillations largely influence the intrinsic GMST The problem of the weak long-range correlation of IMF1 can be solved to a large extent by adding ENSO, IPO, and AMO as forecast precursors to the IMF1 prediction. After completing one prediction, the training set is updated again. It is referred to the above model as E-LSTM (ENSO) [6]. Finally, it fits the residuals by the exponential smoothing model and adds IMF1-5 to the residuals to obtain the prediction results.

DOI: 10.54254/2755-2721/58/20240725

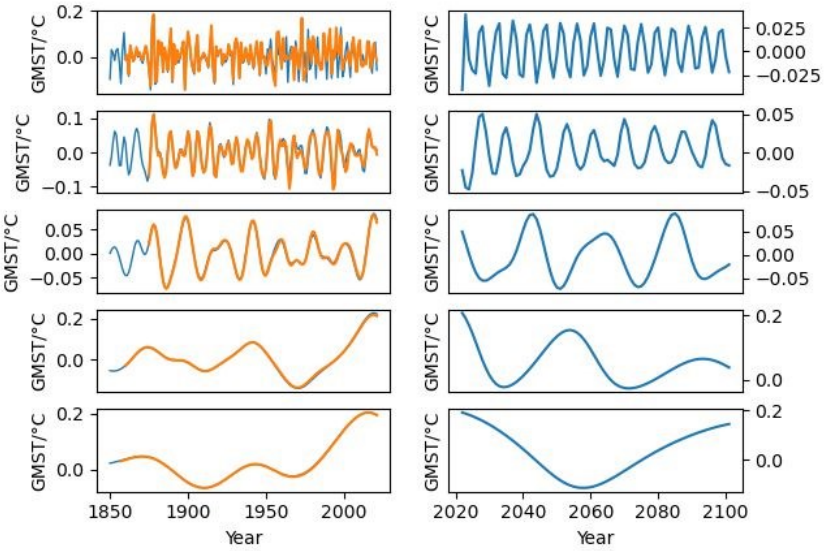


Figure 5. Prediction and fitting results of the LSTM model for each suborder (IMF1-5) after EEMD decomposition.

4. Analysis of Model Accuracy

For a total of 172 data samples of the ARIMA model and E-LSTM model from 1850-2022, the first 100 data samples are used as the training set the last 72 data samples are used as the prediction set, and the goodness-of-fit R and MSE are used as evaluation indicators to analyze the accuracy of the prediction model.

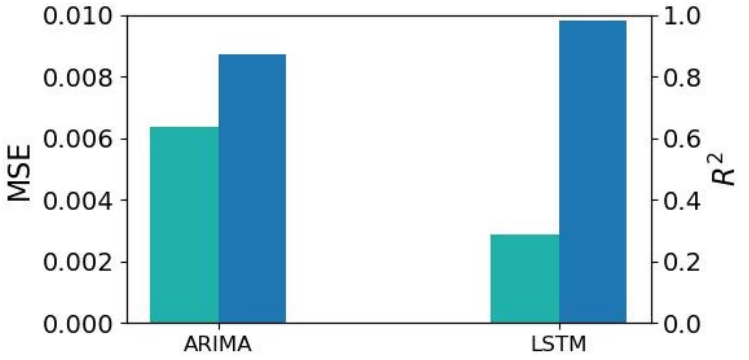


Figure 6. Fitting forecast indicators.

Figure 6 reflects Fitting and forecast indicators. It can be seen that the individual errors of the LSTM model with simultaneous ENSO are significantly lower than those of the other models. Considered together, the E-LSTM model is considered to be more accurate.

5. Results & Discussion

From the results obtained from the above models, it is agreed that the global temperature.

Agree that the increase in global temperature in March 2022 leads to a greater increase than that observed during any previous 10-year period.

Disagree with the prediction that the global average temperature at the observation point will reach 20 ℃ in 2050 or 2100.

DOI: 10.54254/2755-2721/58/20240725

The global average temperature is predicted to reach 14. 15 ℃ in 2050 and 14.51 ℃ in 2100 by ARIMA. The LSTM predicts that the global average temperature will reach 14.30 ℃ in 2050 and 15.60 ℃ in 2100.

The E-LSTM model predicts around 2200 years, and ARIMA predicts around 2860 years when the global average temperature reaches 20 ℃.

By comparing the goodness-of-fit R and MSE, E-LSTM is considered to be more accurate.

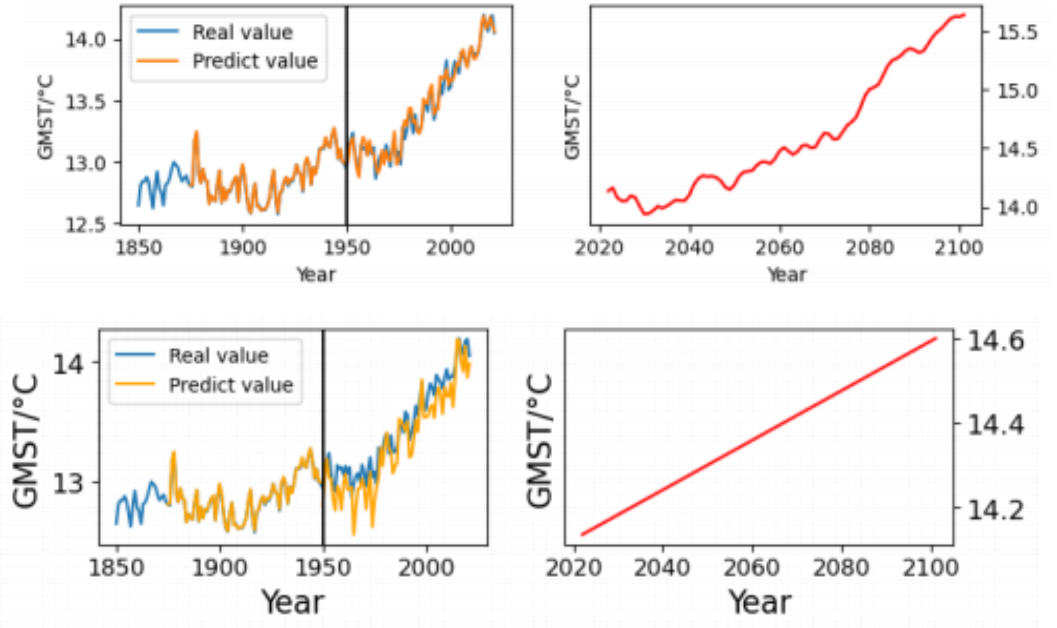


Figure 7. Fitting forecast results (The first line is the result of E-LSTM; The second line is the result of ARIMA).

6. Model and Solution of Problem 2

6.1. Data processing

It is obvious that there are missing data in the dataset, for example, in the data given by the APMCM organizing committee. Many cities are missing temperature data for one two or three consecutive years, and a combination of LOCF (Last Observation Carried Forward) and NOCB (Next Observation Carried Backward) is used to fill in the missing data.

The steps of the algorithm to fill in the missing data are given below:

Step 1: Record the value of the previous year if it is not empty.

Step 2: Record the value of the next year if it is not empty.

Step 3: If the value of this year is empty, fill in the average value of the recorded data.

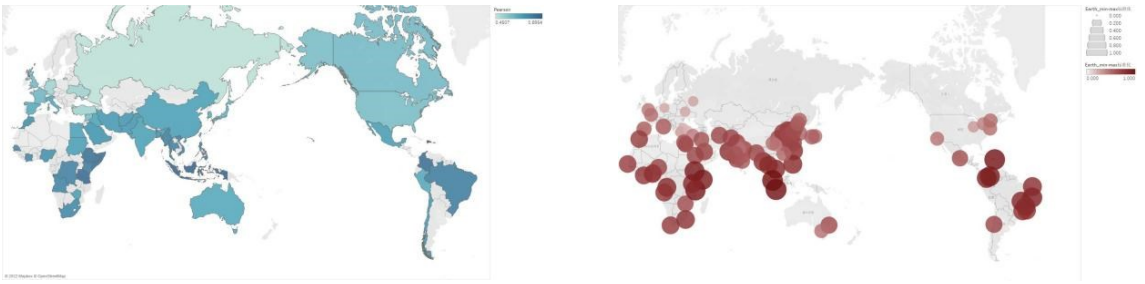
6.2. Pearson Correlation Model

The Pearson correlation [7] model is a model for correlation analysis based on normal continuous distribution data, and its model expression is as follows the following equation is shown:

$$\rho_{x,y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \tag{2}$$

where X, and Y are the two factors to be analyzed and E denotes the mathematical expectation.

DOI: 10.54254/2755-2721/58/20240725



(a) Correlation between global temperature and country (left)

(b) Correlation between global temperature and city (right)

Figure 8. Correlation analysis results.

The P-Values of the global average temperature and the Pearson correlation coefficient of each city are all much less than 0.05, indicating that the model is statistically significant, while 57.4% of the cities have correlation coefficients greater than 0.7, each average correlation coefficient has 0.7064, and the median has 0.7105. Therefore, it is believed that there is a strong correlation between temperature change in each city and global temperature warming.

This paper analyzes the correlation coefficients obtained with the geographical location of each city, Figure 11 shows that the correlation is generally higher for cities near the equator, while the correlation is relatively lower for cities at high and middle latitudes, and the correlation tends to be weaker the higher the latitude, so it can be believed that the closer the region to the equator, the stronger the correlation with global warming, and the farther the region from the equator, the weaker the correlation with global warming.

6.3. Grey Relational Analysis [8]

12 representative disaster indicators were selected in the past 50 years: Mass movement (dry), Inspect infection, Volcanic activity, Drought, Epidemic, Earthquake, Wildfire, Landslide, Extreme temperature, Storm, Flood, and make a grey correlation analysis with the global average temperature data.

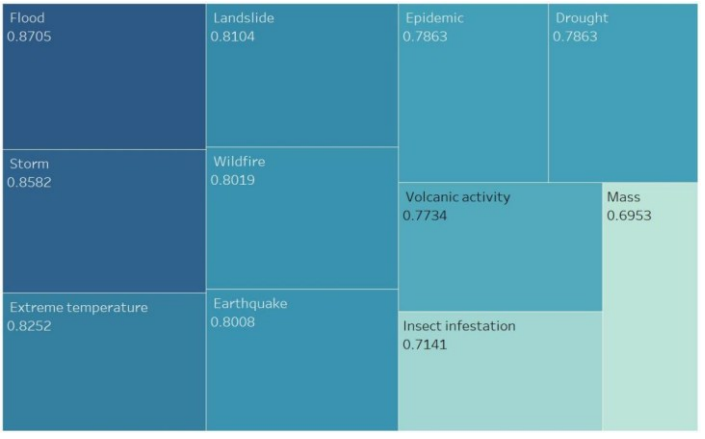


Figure 9. Grey relational analysis results.

In Figure 12, the results showed that Flood, Storm, and Extreme temperatures, had a high correlation with the global average temperature.

6.4. Random Forest Regression

Random forest [9] classification is also a model that integrates multiple CART trees [10] for classification through the idea of integration learning. However, the CART tree here is a classification

DOI: 10.54254/2755-2721/58/20240725

tree, and the evaluation criterion used is different from regression, using the Gini coefficient [11], which is expressed as follows:

$$\text{gini}(p) = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2 \tag{3}$$

where denotes the probability that the selected sample belongs to category k (1 − denotes the probability of being misclassified). Thus, the smaller the Gini coefficient, the smaller the probability of the sample being misclassified, and the more reasonable the division.

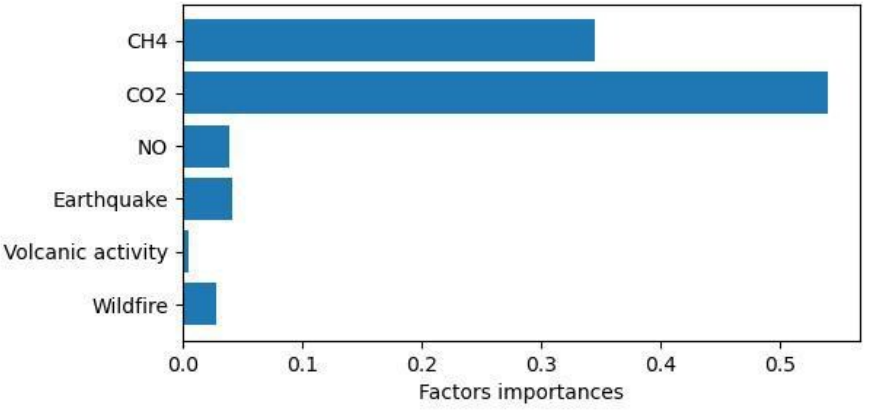


Figure 10. Random Forest regression result.

Figure 10 showed that CO2 concentration and CH4 concentration had significant effects on global average temperature.

6.5. Possible Measures

From the conclusion of 2.c it can be seen that compared to natural disasters, greenhouse gases such as CO2 and CH4 play a greater role in global warming, so for curbing global warming, the first priority is to reduce greenhouse gas emissions and reduce greenhouse gas emissions in the atmosphere, and from 2.c it can be known that the average temperature of the region near the equator has a strong correlation with the global average temperature, so reducing greenhouse gases on the basis focuses on detecting the control of greenhouse gas emissions in the area near the equator.

To reduce greenhouse gas emissions, more clean energy should be used instead of traditional energy, reduce the CO2 and CH4 produced by traditional energy production, and strengthen the management of emitted gases so that greenhouse gases can be emitted in a reasonable way. In addition, it is necessary to strengthen afforestation and increase forest area to enhance the natural ability to reduce CO2.

7. Conclusion

The absorption of heat by the ocean has a great impact on global climate change. Observations have found that changes in ocean surface temperature have some oscillatory characteristics, such as interdecadal Pacific oscillation, El Niño phenomenon, La Niña phenomenon, etc. These factors make it more difficult to study global temperature changes.

Complex professional climate models make it difficult for non-specialists to understand and recognize the dynamics of global climate change and to realize the correlation between climate change, extreme weather and global warming.

References

[1] VIII. ReferencesBox, G.E.P., Jenkins, G.M. (1970) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

DOI: 10.54254/2755-2721/58/20240725

[2] Hochreiter, S., Schmidhuber, J. (1997) Long Short-Term Memory. Neural Computation, 9(8): 1735-1780.
[3] Wu, Z. and N.E. Huang (2009) Ensemble empirical mode decomposition: A noise-assisted data analysis method. Advances in Adaptive Data Analysis, 1(1): 1-41.
[4] Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986) Learning representations by back-propagating errors. Nature, 323(6088): 533-536.
[5] Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H. (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 454(1971): 903-995.
[6] LUO Deyang, ZHENG Fei, CHEN Quanliang. 2022. Prediction of Interannual Signal of Global Mean Surface Temperature Based on Deep Learning Approach [J]. Climatic and Environmental Research (in Chinese), 27 (1): 94−104.doi:10.3878/j.issn.1006-9585.2021.2014
[7] Pearson, K. (1895) Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 58: 240–242.
[8] Deng, J.L. (1982) Control problems of grey systems. Systems & Control Letters, 1(5): 288-294.
[9] Breiman, L. (2001) Random forests. Machine Learning, 45(1): 5-32.
[10] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984) Classification and regression trees. Wadsworth International Group, Belmont, California.
[11] Gini, C. (1912) Variability and Mutability, Contribution to the Study of Statistical Distribution and Relatives. Studi Economico-Giuricici della R.