Validity Investigation and Temperature Prediction of Global Warming Based on Random Forest Algorithm

Shikang He *

College of Information Management, Nanjing Agricultural University, Nanjing, China, 210031

* Corresponding Author Email: 14773421806@163.com

Abstract. This study initially examined the validity of the global warming argument through normal distribution test and one-sample t-test, and conducted a data perspective analysis to determine whether the increase of global temperature in March 2022 was larger than any previous 10-year period. Subsequently, based on the data collection and processing of global temperature impact factors and correlation analysis, three prediction models were established: multiple regression prediction model, ARIMA time series prediction model, and random forest-based nonlinear prediction model. These models collaborated to complete a forecast of future global temperatures. After accuracy assessment and sensitivity analysis with confidence intervals ranging from 95% to 105%, the nonlinear prediction model based on the Random Forest algorithm was selected to further study the future global temperature. The results showed that the global average temperature would reach 20 degrees in 2055, 19.96 degrees in 2050, and 20.07 degrees in 2100.

Keywords: Global Warming, Temperature Prediction, Random Forest, Multiple Regression, ARIMA Model.

1. Introduction

In recent years, many places have broken historical records for high temperatures, and the instability of the global climate system has increased, with frequent, widespread, intense and concurrent extreme weather and climate events.

Global warming has moved from the 'behind the scenes' to the 'in front of the stage'. Global warming is a phenomenon associated with nature. It is caused by the continued accumulation of the greenhouse effect [1-4], which leads to an imbalance in the amount of energy absorbed and emitted by the Earth-atmosphere system, and the continued accumulation of energy in the Earth-atmosphere system, which leads to an increase in temperature and global warming [5-8].

Global warming is a major environmental problem facing mankind, with a wide range of influence, long duration, complex constraints and serious consequences, which is related to the social and economic life of mankind and has attracted the high attention of scholars, the public and governments worldwide. To investigate the impact of global warming, predict global temperature changes, formulate reasonable and efficient development strategies based on scientific evidence, and reduce and control greenhouse gas emissions is an urgent issue to achieve sustainable human development [9].

Investigate the validity of global warming claims, determine whether the increase of global temperature in March 2022 resulted in a larger increase than observed over any previous 10-year period, develop multiple models to complete descriptions of past global temperature levels and projections of future global temperatures, conduct point-specific analyses, and complete assessments of model accuracy and precision [10].

First of all, through the collection of relevant data and hypothesis testing methods, this study compares and explores the extent of global temperature rise in the past ten years, and verifies the relevant claims, and then selects the relevant influencing factors according to the statistical relativity. the establishment of three prediction models is completed on the basis of scientific statistical analysis. Finally, through the sensitivity and error analysis, the model is evaluated and selected, and a conclusion is drawn.

2. Data acquisition and preprocessing

2.1. Data Desceiption

In order to analyze the relationship between various factors and global climate change, relevant data sets need to be collected. Due to the different statistical departments of different elements, the following table I shows the sources of data sets used in this chapter and the meanings of variables.

Tab. 1 Dataset source

Global Average Temperature (Celsius)	Berkeley Earth
Degree of urbanization	Ourworldindata website
Ozone hole area	Ourworldindata website
Sulfur dioxide Forest cover rate	Ourworldindata website
Built-up area	Ourworldindata website
Grazing area	Ourworldindata website
Cultivated area	Ourworldindata website
Annual carbon dioxide emissions	Ourworldindata website

2.2. Data Cleaning

2.2.1 Data missing value processing

Due to the different departments and cycles of data recording, the amount of data in different data sets is not the same. In order to make the data sets of each variable available, the missing values in the data sets should be dealt with accordingly. In this paper, the regression interpolation method based on the serial trend is used to deal with the missing values of the data.

2.2.2 Data standardization processing

Since the units and orders of magnitude of each variable are different, in order to facilitate the construction of a multivariate regression model, the data of each variable can be standardized. In this paper, z-score standardization is used to process the data. It can be represented by the following formula:

$$z = (x - \mu)/\ \sigma \tag{1}$$

x is a specific index of the variable, μ is the mean of the data sequence, and σ is the standard deviation of the data sequence.

2.2.3 Factor correlation analysis chart

The degree of correlation between each factor and temperature is visualized, and the correlation coefficient is drawn as the above heat map. It can be seen from the figure 1 that the above six influencing factors have high correlation and can be used as predictive temperature factors.

	Temperature	Population	CO2	Urbanization	Land area	Forest	SO2
Temperature	1	0.785	0.761	0.847	0.847	-0.897	0.724
Population	0.785	1	0.994	0.992	0.787	-0.954	-0.244
CO2	0.761	0.994	1	0.984	0.731	-0.932	-0.331
Urbanization	0.847	0.992	0.984	1	0.836	-0.98	-0.184
Land area	0.847	0.787	0.731	0.836	1	-0.914	0.367
Forest	-0.897	-0.954	-0.932	-0.98	-0.914	1	0.032
SO2	0.724	-0.244	-0.331	-0.184	0.367	0.032	1

Fig. 1 Correlation coefficient heat map

2.2.4 Non-numeric variable processing

Digitize the non-numeric features, split the date into year and month, split the latitude into positive and negative, with 0 as the boundary, the north latitude is a positive number, and the south latitude is a negative number. Split the longitude into positive and negative, with 0 as the boundary, the east longitude is a positive number, and the west longitude is a negative number.

3. Models and Solutions

3.1. The technical route of this research

The technical route of this research is shown in figure 2.

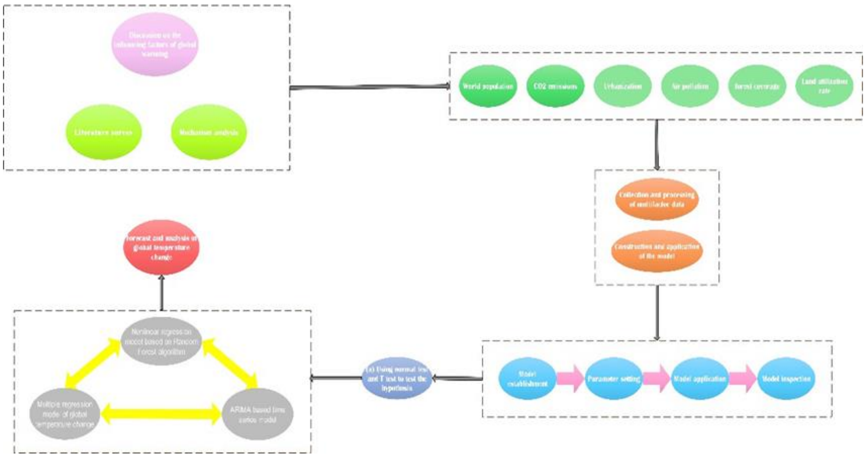

Fig. 2 Structure diagram.

3.2. Verification and judgment

The sequence diagram of the global average temperature from March 2013 to March 2022 by collecting data is as follows (Fig. 3).

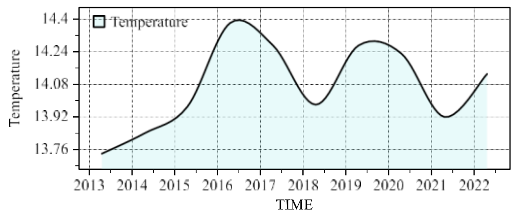

Fig. 3 Series map of global average temperature from March 2013 to March 2022

It can be seen from the Fig. 4 that the global average temperature in March every year has fluctuated and increased in the past ten years. The trendline is:

$$y = 0.0288x - 44.015 \tag{2}$$

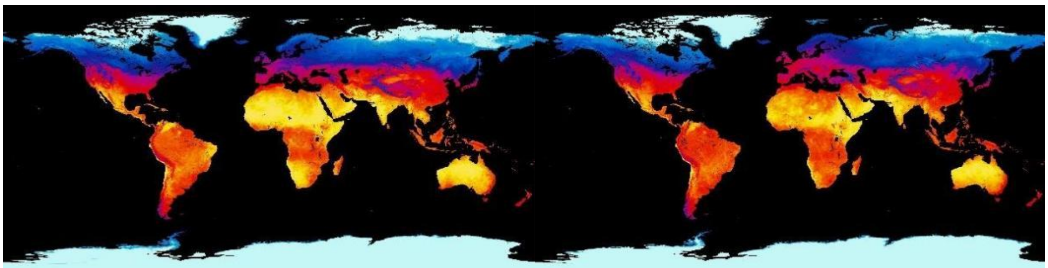

Fig. 4 Comparison of global temperature heat maps between March 2013 and March 2022

As can be seen from the above figure, the global temperature will rise to a certain extent from March 2013 to March 2022. Make a first-order difference of the global average temperature in March every year in the past ten years, and the results are as follows (Fig. 5). It can be seen that the global average temperature has fluctuated and increased in the past ten years, and the growth rate fluctuates around 0.

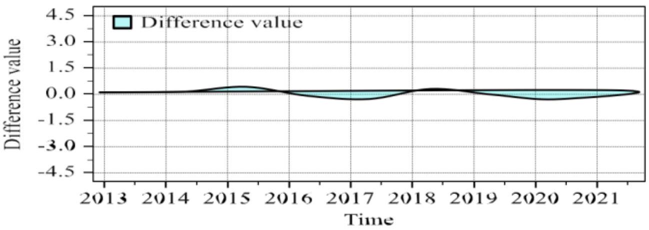

Fig. 5 The first-order difference map of the global average temperature in March each year in the past ten years

Perform a hypothesis test on the first-order difference of the global average temperature in the past ten years and take the original hypothesis H0: the temperature difference between 2021 and 2022 is less than or equal to the temperature difference in the past ten years. Considering the small amount of data, the normal distribution test The Shapiro-Wilk test is suitable. Carry out the Shapiro-Wilk test on the temperature difference in the past ten years to verify whether the data obey the normal distribution. After calculating the Wilk statistics, the corresponding p value can be obtained. Compare the p value with 0.05, if If it is less than 0.05, the null hypothesis can be rejected, that is, the data does not obey the normal distribution, otherwise the null hypothesis cannot be rejected, that is, the data obeys the normal distribution. The results of the Shapiro-Wilk test are as follows (Table II).

Tab. 2 Normality Test

		Kolmogolov-Sminov a			Shapiro-Wilk	
	statistics	degrees of freedom	significant	statistics	degrees of freedom	significant
difference	0.136	8	0.200*	0.954	8	0.756

Note: *. True lower limit of significance.

Riley's significance correction: It can be seen from the table that the p-value (that is, significance) of this group of data is 0.756, which is greater than 0.05, indicating that this group of data obeys a normal distribution. If the data satisfies a normal distribution, a one-sample t-test can be performed on the sample, and the p-value of the t-test can be used to determine whether the null hypothesis H0 can be rejected.

The results of the one-sample t-test are as follows (Table III). It can be seen from the table that the p value of this group of data is 0.08, which is less than 0.1, indicating that the null hypothesis H0 can be rejected at the 90% confidence level, and the alternative hypothesis H1 can be accepted, that is, the temperature difference between 2021 and 2022 is greater than the temperature in the past ten years difference.

Our team agrees that the increase in global temperatures in March 2022 resulted in a larger increase than observed over the past 10 years.

Tab. 3 One-Sample Tests

	test value = 0.21					
	t	**degrees of freedom**	**significant (double tail)**	**average value difference**	**95% confidence interval for the difference**	
					lower limit	upper limit
difference	2.049	7	0.080	0.18750	-0.0289	0.4039

3.3. Establishment and solution of prediction model

3.3.1 Multiple regression prediction

Step 1 Model establishment

Stationary analysis of global climate change: Plotting the year-by-year data of global temperature into the following sequence diagram, it can be seen that the sequence has a gradual upward trend, which shows that the trend of global warming does exist (Fig. 6).

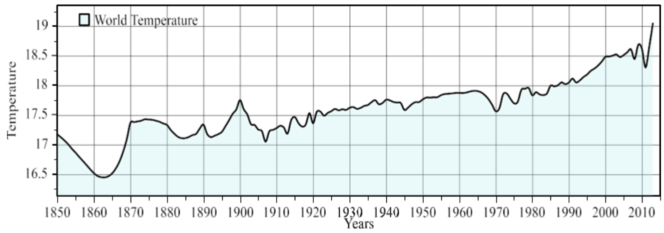

Fig. 6 Year-by-year data series of global temperature

After the yearly data of global temperature are processed by first-order difference, the sequence obtained is relatively stable (Fig. 7).

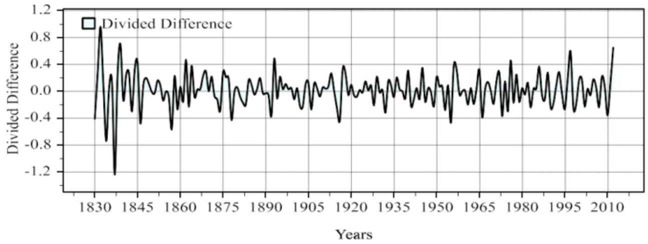

Fig. 7 Sequence diagram of yearly data of global temperature after first-order difference processing

(1)Model construction

Combined with practical considerations and according to relevant research [2], the main factors affecting global climate change include world population, atmospheric carbon dioxide emissions, forest coverage, urbanization rate, land use area, sulfur dioxide emissions, etc.

Establish a multiple linear regression model between each factor and global temperature and use the training set and test set to evaluate the accuracy of the model, and finally use the constructed model to predict the global temperature in 2050 and 2100.

(2)Build multiple linear models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon \tag{3}$$

Multiple linear regression model establishment

$$y = 17.439 + 0.400x_1 + 0.461x_2 + 0.180x_3 + 0.900x_4 + 1.357x_5 + 0.004x_6 \tag{4}$$

Tab. 4 Model Summary

Model	R	R square	Adjusted R square	Standard Estimate Error
	0.925[a]	0.855	0.850	0.25703

The model summary shows that the R-square of the model is 0.855, the adjusted R-square is 0.850, and the overall fitting effect is good (Table IV).

Hypothesis is raised in F-test: all coefficients are 0. From the ANOVA table (see appendix), it can be seen that the p value of the F test is <0.001, and the null hypothesis is rejected at the significance level of 0.05, and the test is passed.

Step 2 Forecast and Analysis

Model prediction results and testing: The multiple linear regression model is used to fit and predict the global temperature. The fitting results of the model are shown in the figure below. It can be seen that the model fits the historical data well, and the change trend is the same as the real trend (Fig. 8).

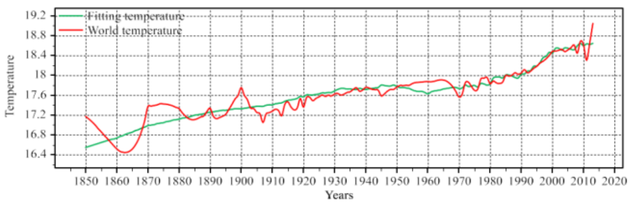

Fig. 8 Temperature Forecast Time Series

The forecast results of the global average temperature for the future to 2100 are as follows (Fig. 9):

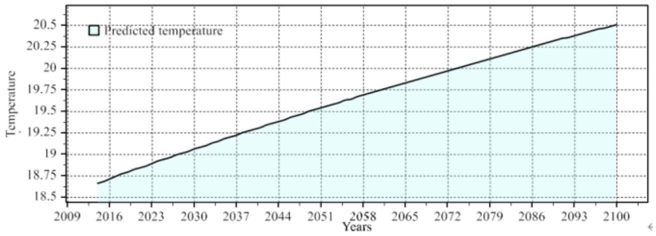

Fig. 9 Average temperature forecast results.

Conclusion: From the temperature curve obtained by the multiple linear regression model, it can be concluded that the trend of global climate change in the future is that the temperature will continue to rise, and the world will continue to warm.

3.3.2 ARIMA time series forecasting

The ARIMA model is Autoregressive Integrated Moving Average Model, which is a model established for non-stationary time series.

Step 1 Model building

From the establishment process of the establishment of the multiple regression model, it can be seen that the stability of the first-order difference sequence diagram has been significantly improved, which is conducive to the establishment of the ARIMA model. After the analysis of stationarity, periodicity, autocorrelation and partial autocorrelation, it is found that the effect of the model is the best when the ARIMA (0, 1, 5) model is used for fitting and forecasting (Table V). The R-square of the model is 0.881, indicating that the results of the model have high credibility.

Step 2 Forecast and Analysis

The residual ACF (autocorrelation coefficient) and PACF (partial autocorrelation coefficient) plots of the model are as follows (Fig. 10).

Tab. 5 Model Statistics

Model	predictor variable	Model Fit Statistics		Young Box Q (18)			Outliers
		Stationary R square	Rsquare	statistics	DF	significance	
Air temperature - model	0	0.218	0.881	10.496	15	0.787	0

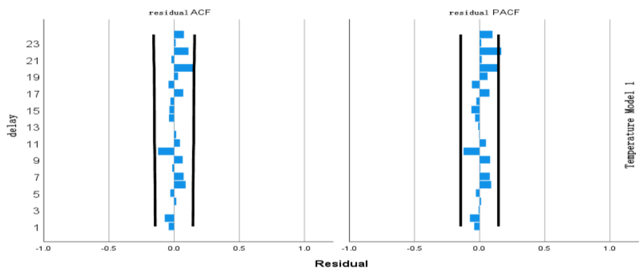

Fig. 10 Residuals plot of time series model based on global temperature.

The model's fitting and forecasting graph of global temperature is as follows (Fig. 11). It can be seen from the figure that the model fits the historical data better, and the change trend is consistent with the real trend.

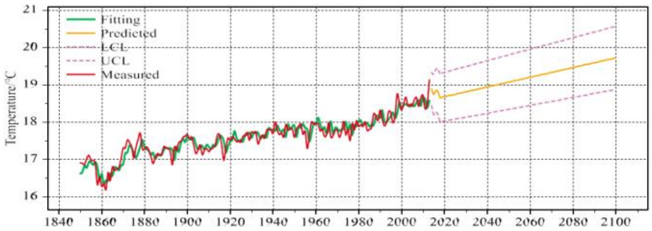

Fig. 11 Fitting and forecasting plot of time series model based on global temperature.

3.3.3 Nonlinear prediction based on random forest algorithm

Step 1 Model building

Random forest regression is a model that integrates multiple CART trees for regression prediction through the idea of ensemble learning. In this paper, the CART regression tree is used to build the model. Based on the principle of minimum mean square error, each factor leaf node (division point) s is divided for each climate influencing factor. For any division factor A, the corresponding arbitrary division point s, the parent node splits into two child nodes (hypothesis), the data sets of the two nodes are D1 and D2, find the corresponding factors that minimize the mean square error of the respective sets of D1 and D2, and minimize the sum of the mean square errors of D1 and D2 and factor value as the dividing point.

The expression is:

$$\min_{A,s}[\min_{c_1}\sum_{x_i\in D_1(A,s)}(y_i-c_1)^2+\min\sum_{x_i\in D_2(A,s)}(y_i-c_2)^2] \tag{5}$$

In the formula, c1 represents the mean value of the first node, and c2 represents the mean value of the second node. The above formula is used as the principle to determine the factor division points, and finally establish a climate prediction regression model based on each factor.

Step 2 Random Forest Model Parameter Tuning

From the optimal parameter training diagram (see appendix), it can be seen that when the Number of Leaves is 5, the parameters are optimal, and the mean square error is 0.01.

Step 3 Evaluation of factor contribution rate(Fig. 12)

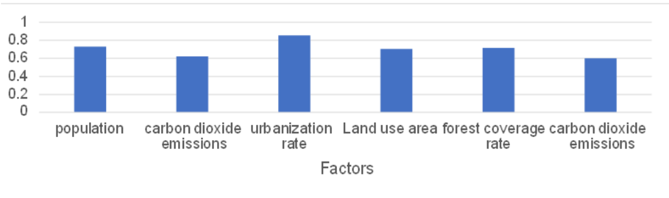

Fig. 12 Importance of climate-influencing factors

Step 4 random forest prediction effect evaluation

The prediction effect of the random forest was tested on the test set, and the results showed that the adjusted R^2 value of the prediction reached 0.9586, and the prediction effect was good(Fig. 13 and 14).

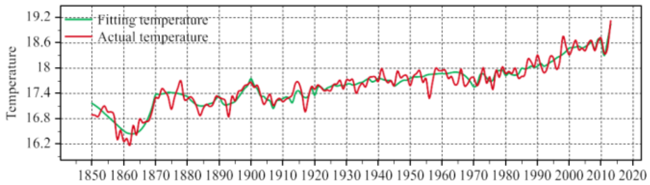

Fig. 13 Random Forest fitting data

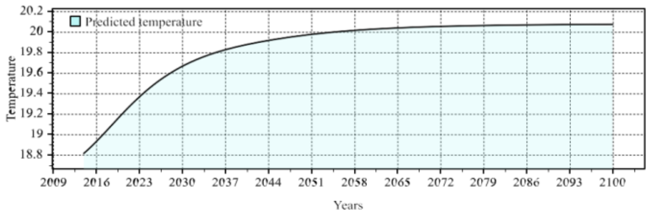

Fig. 14 Global Average Temperature Forecast

3.4. Conclusions of the model

3.4.1 Multiple regression prediction model

By 2050, the global average temperature is predicted to be 19.51 degrees, and by 2100, the global average temperature is predicted to be 20.51 degrees, that is, the model does not believe that the global average temperature will reach 20 degrees in 2050 or 2100, and the average temperature will reach 20 degrees The time forecast is 2074.

3.4.2 ARIMA time series model

By 2050, the global average temperature is predicted to be 19.06 degrees, and by 2100, the global average temperature is predicted to be 19.71 degrees, that is, the model does not believe that the global average temperature will reach 20 degrees in 2050 or 2100, and the average temperature will reach 20 degrees The time forecast is 2122.

3.4.3 Nonlinear prediction model based on random forest algorithm

By 2050, the global average temperature is predicted to be 19.96 degrees, and by 2100, the global average temperature is predicted to be 20.07 degrees, that is, the model believes that the global average temperature will reach 20 degrees in 2050 or 2100, and the average temperature will reach 20 degrees The time forecast is 2055.

3.5. Model Precision and Accuracy

It can be seen from Table VI that when the six influencing factors are used for modeling, the absolute error and relative error of the five-year independent sample forecast by the multiple regression method are 0.148°C and 0.80% respectively; Under these conditions, the absolute error and relative error of the ARIMA time series forecast for the six-year independent sample forecast are 0.016°C and 0.87%, respectively, and the absolute error and relative error of the six-year independent sample forecast for the nonlinear forecast based on random forest are 0.081579°C and 0.44%

Therefore, it can be concluded that in this question, the nonlinear forecasting ability based on random forest is stronger than the multiple regression forecasting model and the ARIMA time series forecasting model.

Tab. 6 Forecasts and errors of the three models

years	reality	Multiple Regression Forecasting			ARIMA time series forecasting			Nonlinear Forecasting Based on Random Forest		
	forecast		absolute error	relative error	forecast	absolute error	relative error	forecast	absolute error	relative error
2008	18.41	18.59	0.18	0.98%	18.58	0.17	0.92%	18.43831	0.02831	0.15%
2009	18.63	18.64	0.01	0.05%	18.42	0.21	1.13%	18.68301	0.05301	0.28%
2010	18.71	18.59	0.12	0.64%	18.64	0.07	0.37%	18.61449	0.09551	0.51%
2011	18.35	18.63	0.28	1.53%	18.6	0.25	1.36%	18.29478	0.055222	0.30%
2012	18.47	18.62	0.15	0.81%	18.37	0.1	0.54%	18.64674	0.17674	0.96%
	Average		0.148	0.80%		0.16	0.87%		0.081579	0.44%

4. Conclusions

This study demonstrated the effectiveness of the random forest algorithm in predicting global temperature changes. The results provide valuable insights for further research on global warming and contribute to the development of strategies to mitigate the impact of climate change. With the increasing accuracy of prediction models, policymakers can make more informed decisions to ensure a sustainable future.

References

[1] Azadeh R, Ali T, Leila S. Light pre-conditioning at the seedling stage affected root growth behavior and enhanced the thermotolerance of Petunia × hybrida [J]. Scientia Horticulturae, 2024, 325.

[2] Jianying L, Yang C, Jiangyu M, et al. Increase in optimal configuration of 25–60-day atmospheric circulations for Yangtze heavy rainfall under global warming background [J]. Weather and Climate Extremes, 2023, 42.

[3] Uygun S B, Ibrahim D. A study on comparative environmental impact assessment of thermochemical cycles and steam methane reforming processes for hydrogen production processes [J]. Computers and Chemical Engineering, 2024, 180.

[4] Lewis M J, Ariane P S, Muhammad A N, et al. Synergy between industry and agriculture: Techno-economic and life cycle assessments of waste recovery for crop growth in glasshouses[J]. Journal of Cleaner Production, 2023, 432.

[5] Pingping H, Feng Z, Bailing Z, et al. Active microeukaryotes hold clues of effects of global warming on benthic diversity and connectivity in the coastal sediments [J]. Ecological Indicators, 2024, 158.

[6] Ju-Seok H, Hyo-Seok P, Eui-Seok C. Projections of central Arctic summer sea surface temperatures in CMIP6 [J]. Environmental Research Letters, 2023, 18 (12).

[7] Chen F, Ya S, Qianting Z, et al. Responses of greenhouse gas emissions to aeration coupled with functional membrane during industrial-scale composting of dairy manure: Insights into bacterial community composition and function [J]. Bioresource Technology, 2024, 393:130079-130079.

[8] Faisal M, Guangshuai W, Amin S A, et al. Optimizing irrigation management sustained grain yield, crop water productivity, and mitigated greenhouse gas emissions from the winter wheat field in North China Plain [J]. Agricultural Water Management, 2023,290.

[9] Andrew L, Bilainu O, Bilal P. Life cycle assessment of Power-to-Gas (PtG) technology– Evaluation of system configurations of renewable hydrogen and methane production [J]. Sustainable Energy Technologies and Assessments, 2023, 60.

[10] Mohammad Y, Hussameldin I . A comprehensive review on recent trends in carbon capture, utilization, and storage techniques [J]. Journal of Environmental Chemical Engineering,2023, 11 (6)