Research on Carbon Dioxide Concentration Prediction Based on RNN Model in Deep Learning

Hongxuan Wang, Wenyuan Liang, Siyun Liang*, Bohao Chen

The High School Affiliated to Renmin University of China, Beijing, China, 100080

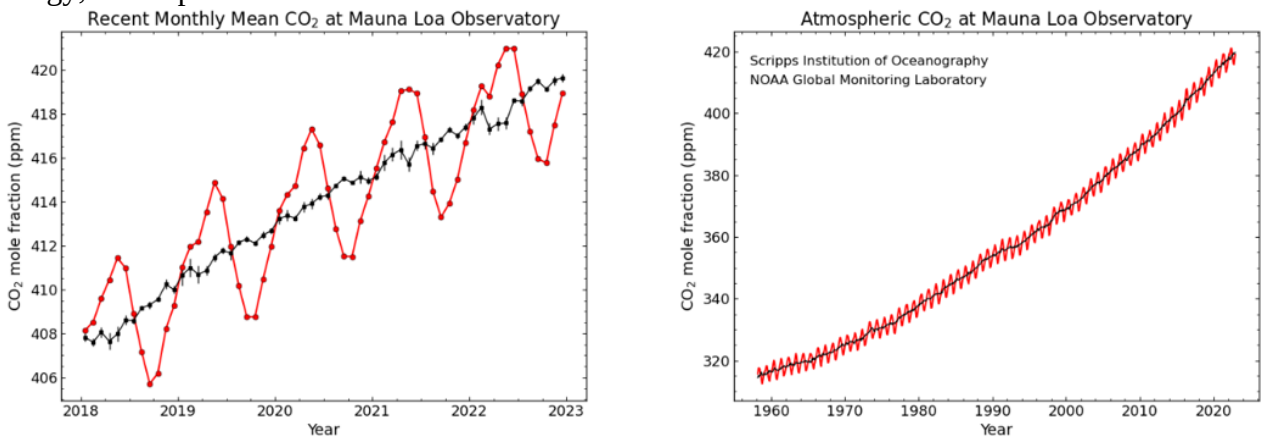
*Corresponding author: liangsy01@outlook.com

Abstract. Predicting the concentration of carbon dioxide and its ensuing effect on ocean temperature relatively accurately is critical in balancing control future economic and industrial development with environmental protection.In this paper, we divided the data into training set (85%) and test set (15%) and selected three typical models, statistical model (Holt-Winters smoothing model), machine learning model (linear regression model), and deep learning model (RNN), and evaluated the models according to the differences between predicted and actual values. The results show that the RNN model has the smallest MAE, MLSE, and MSE and the best prediction accuracy. Secondly, this paper investigates the temperature change and the relationship between carbon dioxide and land-ocean temperature, and obtains the correlation coefficient between carbon dioxide concentration and temperature change as high as 0.961. With a view to provide some reference significance for ecological conservation and sustainable development.

Keywords: RNN, Prediction Model, Deep Learning, Carbon Dioxide.

1. Introduction

Carbon dioxide (CO2) is a significant heat-trapping gas that is produced by the extraction and combustion of fossil fuel, wildfires, and natural processes such as volcanic eruptions. According to the data provided by The National Aeronautics and Space Administration, human activities have increased atmospheric CO2 by approximately 50% since the beginning of industrial times (in the 18th century). Since modern recordkeeping began in 1955, the internal heat of the ocean has increased due to the ocean's 90 percent contribution to global warming. The surface holds the majority of the extra energy, at a depth of 0 to 700 meters.



(a)Recent global monthly means (b)Global monthly means since 1980

Figure 1. Monthly mean carbon dioxide globally averaged over marine surface sites

Figure 1 shows monthly mean carbon dioxide globally averaged over marine surface sites [1]. The changes cause long-term effects on marine biodiversity, as well as the lives and livelihoods of coastal communities and beyond-including approximately 680 million people who live in low-lying coastal areas, nearly 2 billion who live in half of the world's coastal megacities, nearly half of the world's population (3.3 billion) who rely on fish for protein, and nearly 60 million people who work in fisheries and the aquaculture sector globally [2]. Predicting the concentration of carbon dioxide and

its ensuing effect on ocean temperature relatively accurately is critical in balancing control future economic and industrial development with environmental protection.

2. Acquisition of data and assumptions

The data used in this paper were obtained from question B of the HIMCM, a mathematical modeling competition for middle school students in the United States in 2022. In order to facilitate the analysis of the problem, the following assumptions are made about the data used in this paper.

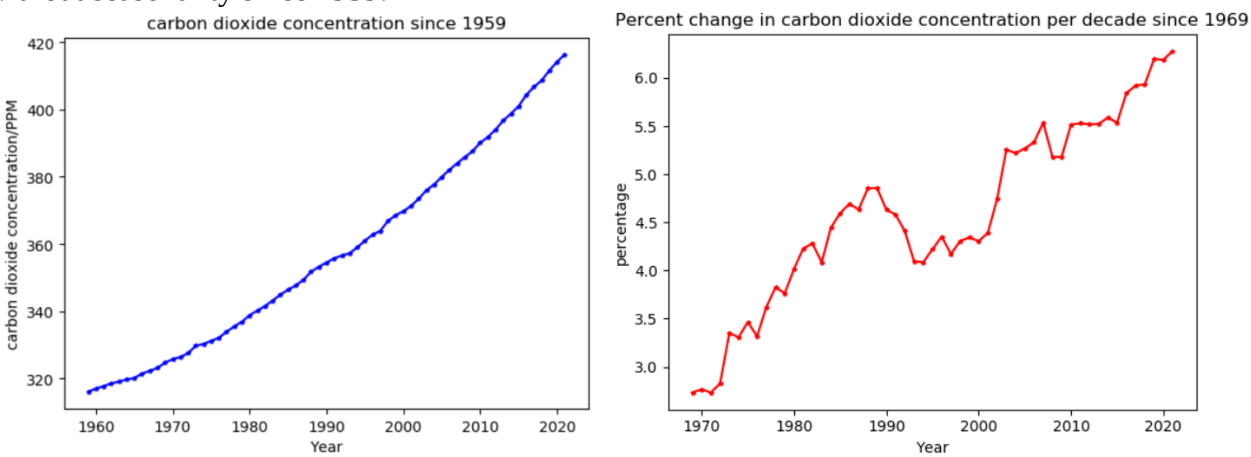
Assumption 1: The data provided is reliable and free from errors. Because forecasting is based on historical data, data is very important.

Assumption 2: People and governments do not take special measures to control carbon dioxide emissions. Our predictions based on historical data on the assumption that future conditions are repetitions of present conditions.

Assumption 3: The CO_2 concentration and temperature change have connection with time(year). It depends on time so we can do a fit test.

3. The changes in carbon dioxide concentration

The dataset used in the work has an increasing trend as shown in Figure 2(a). From Figure 2(a), the trend of carbon dioxide concentration can be observed. Carbon dioxide levels have been rising without seasonality since 1959.



(a)CO_2 concentration since 1959 (b)Percent change in CO_2 concentration per decade since 1969

Figure 2. Change of carbon dioxide concentration

In order to find if the March 2004 increase of CO_2 resulted in a larger increase than observed over any previous 10 - year period or not. The percent change in carbon dioxide concentration per decade since 1969 should be calculated. We use C_{Ci} to represent CO_2 concentration in year i, C_{Ci-10} to represent the CO_2 concentration ten years before year i, use P_{Ci} to represent Percent change in CO_2 concentration per decade in year i. The percent change of CO_2 concentration is calculated by the following formula.

$$P_{Ci} = (C_{Ci} - C_{Ci-10})/C_{Ci-10} \tag{1}$$

The result of percent change of CO_2 concentration is shown in Figure 2(b). From Figure 2(b), we can the trend of percentage is growing, despite the percentage of some years become small. It is obvious that the March 2004 increase of CO_2 not resulted in a larger increase than observed over any previous 10-year period, the 2021 is the largest increase.

4. Methodology introduction and model overview

4.1. The establishment of RNN model

In order to process time-related information, Elman et al. proposed Recurrent Neural Network (RNN) in 1990, which has achieved good performance in natural language processing and computer vision, and has become a widely used deep learning algorithm. It is a good choice to choose RNN model to process the time series problem[3]. For Figure 3 and Figure4, x_t is the input data at time t, h_t is the hidden state at time *t*, h_t is the output at time *t*, and *u*, *w*, *v* are the shared parameters of neurons.

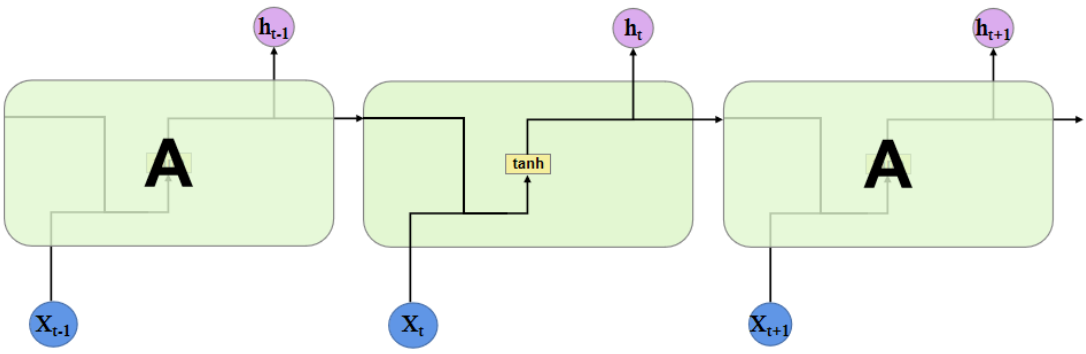


Figure 3. RNN network structure

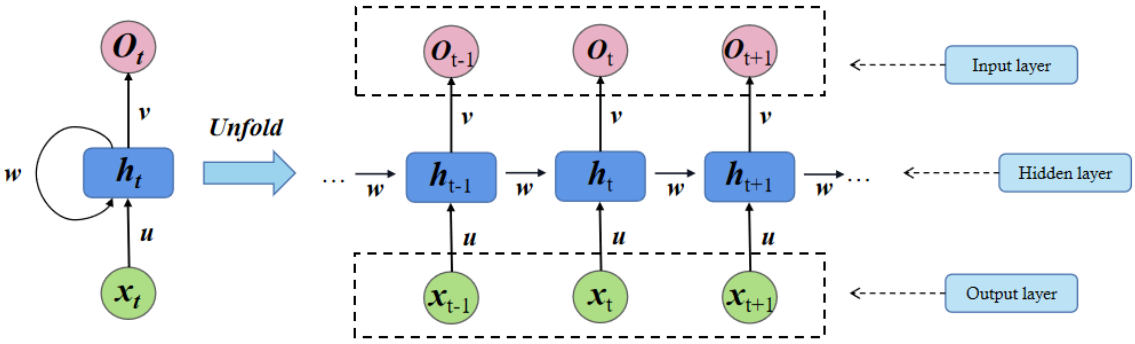


Figure 4. Expansion of recurrent neural network

The recurrent neural unit in Figure 4 can be expanded according to BPTT.

$\{x_1, \dots, x_{t-1}, x_t, x_{t+1}, \dots\}$ is the time series of input, $\{h_1, \dots, h_{t-1}, h_t, h_{t+1}, \dots\}$ is the time state sequence output by the hidden layer, , $\{o_1, \dots, o_{t-1}, o_t, o_{t+1}, \dots\}$ is the time series of output information, u, w, v are neurons Shared parameters[4].

$$h_t = f(w \cdot h_{t-1} + u \cdot x_t) \tag{2}$$

$$O_t = f(v \cdot h_t) \tag{3}$$

f, g is the activation function of the neuron.

$$o_t = g(v \cdot f(w \cdot h_{t-1} + u \cdot x_t)) \tag{4}$$

It can be further written as:

$$o_t = g\left(v \cdot f\left(u \cdot x_t + w \cdot f(u \cdot x_{t-1} + wf(u_{x-2} + \cdots))\right)\right) \tag{5}$$

It can be seen from the above formula that the output of the information at the current moment is related to the input of the information at the previous moment.The results of the RNN model are shown in Figure 5. From the Figure 5, it can be seen that the predicted Holt-Winters exponential smoothing model is accurate.

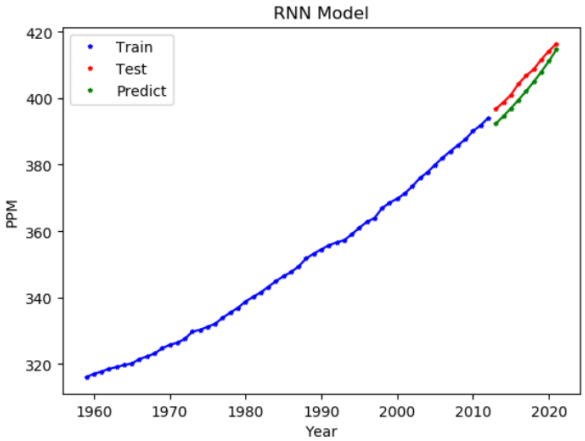


Figure 5. The result of RNN model

4.2. Linear regression model

Linear regression is a machine learning model to solve regression problems. This model solves the problem by assuming a linear relationship between the given input attributes and the output, as shown in following.

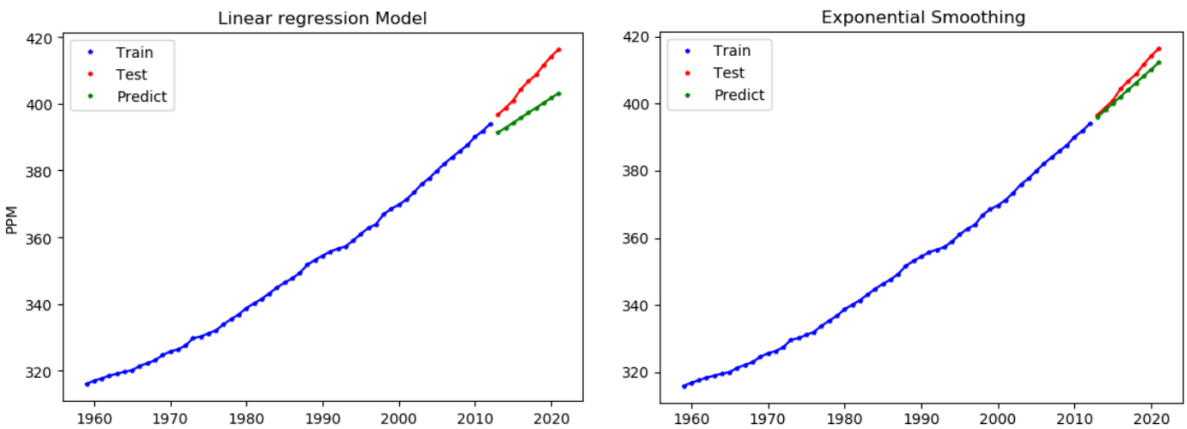
Target output = Input 1 × weight 1 + Input 2× weight 2 …. Input n × weight n + Bias

Here *n* is a total number of attribute/input features. Weights are also called regression coefficients learned while training the model based on train data. In this case, the input data is the year, the output data is carbon dioxide concentration ,it has only one variable[5].

4.3. Holt-Winters exponential smoothing Model

In early 1957, Holt extended the simple exponential smoothing method to make it possible to predict trended data. Secondary exponential smoothing takes into account the baseline and trend of the series, and triple exponential smoothing introduces seasonal components to consider the periodic pattern of time series.Seasonality (periodicity) means that a sequence has a certain repetitive pattern in every fixed time interval, and let represent the length of the cycle.

The Holt-Winters exponential smoothing method has two different seasonal components, the additive model and the multiplicative model[6]. The additive model is usually chosen when seasonal variation remains roughly constant in the time series. The multiplicative model is usually chosen when the seasonal variation varies proportionally to the levels of the time series. Carbon dioxide levels have been rising regardless of season, so the additive model in this case was be chosen. The results of linear regression model and Holt-Winters exponential smoothing model can be shown in Figure 6 (a) and Figure 6 (b) .



(a)Linear regression model (b)Holt-Winters exponential smoothing model

Figure 6. The result of linear regression model and exponential smoothing

5. Evaluate the Models

In this section, this research has discussed the performance metrics used to evaluate the models. These models are used to analyze CO_2 emission, and forecasting, which is a kind of regression problem. There are many evaluation metrics, this research has chosen to evaluate the models. Before using all these metrics, we must know about the residual error, i.e. $(y - \hat{y})$. Here, y and $\hat{y}$ indicate the actual and predicted values. The performance metrics used to evaluate the models are as follows[7].

5.1. Mean squared error

In mean squared error (MSE), first, calculate the squares of the residual error as defined above for each data point and then calculate the average of that [8]. The formula for MSE as follows.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y}_l)^2 \tag{6}$$

5.2. Mean absolute error

Mean absolute error (MAE) is the sum of the absolute residual error. That means it does not matter about negative or positive. The formula for mean absolute error as follows.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \bar{y}_l| \tag{7}$$

5.3. Mean squared log error

In mean squared log error (MSLE), the residual error is calculated with the logarithm of the original and predicated values. It is an extension on mean squared error (MSE), mainly used when predictions have large deviations [9]. The formula for mean absolute log error as follows.

$$MLSE = \frac{1}{n}\sum_{i=1}^{n} \left(\log(y_i) - \log(\bar{y}_l)\right)^2 \tag{8}$$

The results of different models are shown in Table 1. From Table 1, we can find the RNN model is the best model, the Single Exponential Smoothing is the worst model.

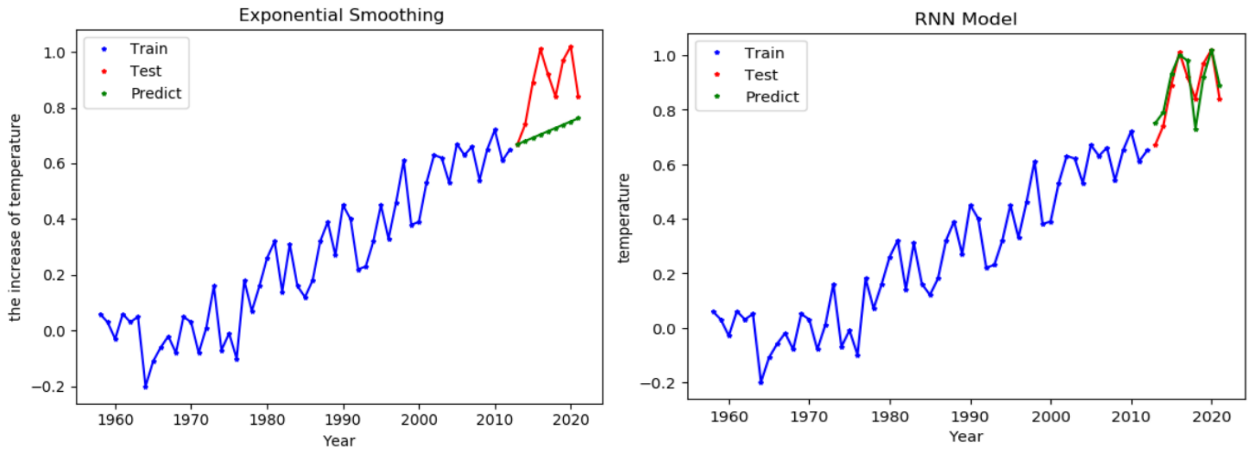
Table.1. The evaluate results of different models

Model Name	MSE	MAE	MLSE
Linear regression	5.18	2.17	0.012
Holt-Winters smoothing	0.74	0.42	0.0012
RNN	0.36	0.23	0.0009

6. Model of land-ocean Temperature change and carbon dioxide concentration

6.1. Model of land-ocean Temperature change

Two models was be chosen to predict the temperature change, the Holt-Winters smoothing model and RNN model. The results are Figure 7(a) and Figure 7(b). Holt-Winters smoothing model has a poor performance in predict the temperature change. The RNN model performance well. Therefore, we finally choose RNN to predict the temperature change. According to RNN model, in 2030, the temperature will change by 1.25 Celsius degrees, in 2042, the temperature will change about 1.5 Celsius degrees, in 2065, the temperature will increase about 2 Celsius degrees.



(a)Holt-Winters exponential smoothing model (b) RNN model

Figure 7. Results of land-ocean Temperature change

6.2. Relationship between CO2 concentration and Temperature change

Correlation coefficient is the first statistical indicator designed by statistician Carl Pearson, it is the amount of linear correlation between research variables, generally expressed by the letter r. Due to the different research objects, there are many ways to define the correlation coefficient, and the Pearson correlation coefficient is more commonly used [10].

The Pearson correlation coefficient is calculated as follows:

$$R = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}} \tag{9}$$

In this case, X represents the value of carbon dioxide concentration, $\bar{X}$ means the average carbon dioxide concentration since 1959, Y represents temperature change, $\bar{Y}$ means the average temperature change since 1959. Get a correlation coefficient as high as 0.961 between CO_2 concentration and temperature change. Therefore, they are highly correlated.

For further research, we build a model linear regression model as we mention above, with 95% confidence bounds, we get the relationship between the CO_2 concentration (x) and temperature change (y).

$$y = 0.01048x - 3.393 \tag{10}$$

In addition, we can see the relationship from the Figure 8, it is obvious they are related, the absolute of residuals is small.

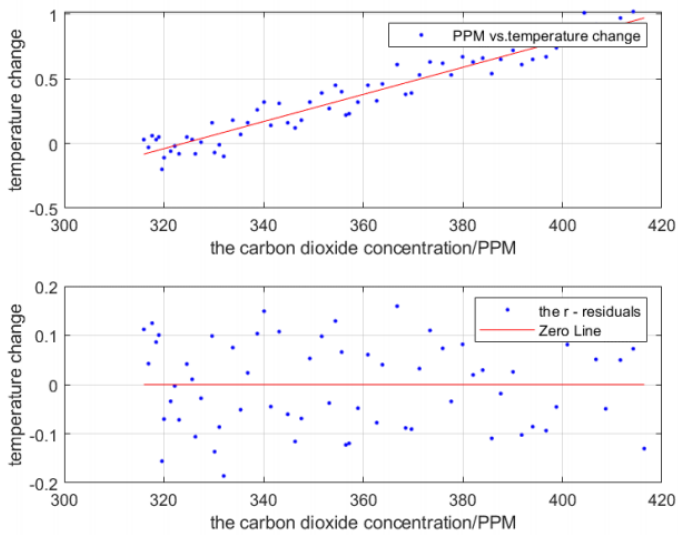


Figure 8. The relationship between CO2 concentration and temperature change

7. Conclusions

The rise in global CO2 levels will lead to global warming, which will have a huge impact on the Earth. In this paper, the data are divided into training set (85%) and test set (15%), and three typical models, statistical model (Holt-Winters smoothing model), machine learning model (linear regression model), and deep learning model (RNN), are selected to evaluate the models based on the difference between predicted and actual values. In this paper, MAE, MLSE, and MSE are selected to evaluate the models. The results show that the RNN model has the best prediction accuracy. Secondly, this paper investigates the temperature variation and the relationship between CO_2 and land-ocean temperature, and gets the correlation coefficient between CO_2 concentration and temperature variation as high as 0.961.

References

[1] Tena-Gago D, Golcarenarenji G, Martinez-Alpiste I, et al. Machine-Learning-Based Carbon Dioxide Concentration Prediction for Hybrid Vehicles[J]. Sensors, 2023, 23(3): 1350.

[2] Caldeira K, Wickett M E. Ocean model predictions of chemistry changes from carbon dioxide emissions to the atmosphere and ocean [J]. Journal of Geophysical Research: Oceans, 2005, 110(C9).

[3] Leonardos E D, Tsujita M J, Grodzinski B. Net Carbon Dioxide Exchange Rates and Predicted Growth Patterns in AlstroemeriaJacqueline'at Varying Irradiances, Carbon Dioxide Concentrations, and Air Temperatures [J]. Journal of the American Society for Horticultural Science, 1994, 119(6): 1265-1275.

[4] Yuan X, Wang J, Deng S, et al. Recent advancements in sustainable upcycling of solid waste into porous carbons for carbon dioxide capture [J]. Renewable and Sustainable Energy Reviews, 2022, 162: 112413.

[5] Safaei-Farouji M, Thanh H V, Dai Z, et al. Exploring the power of machine learning to predict carbon dioxide trapping efficiency in saline aquifers for carbon geological storage project[J]. Journal of Cleaner Production, 2022, 372: 133778.

[6] Kheir A M S, El Baroudy A, Aiad M A, et al. Impacts of rising temperature, carbon dioxide concentration and sea level on wheat production in North Nile delta[J]. Science of The Total Environment, 2019, 651: 3161-3173.

[7] Jumin E, Zaini N, Ahmed A N, et al. Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction[J]. Engineering Applications of Computational Fluid Mechanics, 2020, 14(1): 713-725.

[8] Wihersaari L, Ashton N J, Reinikainen M, et al. Neurofilament light as an outcome predictor after cardiac arrest: a post hoc analysis of the COMACARE trial[J]. Intensive care medicine, 2021, 47: 39-48.

[9] Suter S, Haussener S. Optimizing mesostructured silver catalysts for selective carbon dioxide conversion into fuels[J]. Energy & Environmental Science, 2019, 12(5): 1668-1678.

[10] Wei W, Ramalho O, Malingre L, et al. Machine learning and statistical models for predicting indoor air quality[J]. Indoor Air, 2019, 29(5): 704-726.