Using Holt-Winters Model and Regression Forecast Models to Describe and Predict the Concentration Level of Carbon Dioxide

Ruolin Peng[1]*, Shichen Huang[2,a]

[1]Changjun High School of Changsha City, Changsha, 410023, China
[2]Beijing No.2 middle school international division, Beijing, 100010, China

*2894295090@qq.com; a3605016704@qq.com

Abstract. In this paper, we generally focus on analyzing the concentration of carbon dioxide from the past and predict it in the future. And by using the carbon dioxide concentration data from previous years, we established Holt-winter model, linear regression model and ternary regression model to predict the future trends. Then, we find out that the concentration level of carbon dioxide has a rapid increasing tendency in the future, and in 2050 the concentration will reach 505.23 parts per million (ppm) but not 685 ppm based on the ternary regression model, the most accurate model that was used in this paper. Also, we find out that the concentration of CO2 will reach 740 ppm in 2100. In fact, the safety range of the concentration of carbon dioxide is 400 ppm to 700 ppm. This means that governments must take actions and draw up a plan in order to fight against with global warming, or the global warming will become more and more serious which will lead to more severe ecosystem crisis.

Keywords: ternary linear regression model, Holt-winters model, simple linear regression model, carbon dioxide, global warming

1 Introduction

Global warming becomes more and more serious after industrial revolution. It is a very severe problem when it comes to global warming. The atmospheric CO_2 concentration has risen rapidly since the industrial revolution. Over the last two centuries, it has increased from 280 parts per million (ppm) to the current 410 ppm, which is higher than at any point in the past 800,000 years.[1] If we allow it to go like this, global precipitation would be redistributed, glaciers would melt in decades, and sea level would rise rapidly. Therefore, this will not only destroy the ecological balance, but also threaten people's original living and life seriously. In this case, it is necessary for people to be aware of the severe consequence that global warming has on their life and also those with authority have to work out arrangements as soon as possible to slow down greenhouse gas emissions and reduce the greenhouse effect. But how does government plan scheme if they don't have any idea about what the carbon dioxide emission will going on in the

Z. Ahmad et al. (eds.), *Proceedings of the 2024 5th International Conference on Urban Construction and Management Engineering (ICUCME 2024)*, Advances in Engineering Research 242,
https://doi.org/10.2991/978-94-6463-516-4_3

future? Therefore, in addition to the policy, it is also a vital aspect for experts to establish certain mathematical models in order to predict the tendency of global warming and the emission of the harmful gases, especially carbon dioxide.

This paper will exactly focus on analyzing four problems that are listed below. First, using the data set to analyze the change of CO_2 in the past. Nowadays, carbon dioxide (CO_2) produce in the atmosphere becomes larger than before. Prior to the Industrial Revolution, carbon dioxide (CO_2) in the atmosphere was consistently around 280 ppm. The concentration of CO_2 in the atmosphere reached 377.7 ppm, which is a huge amount of increment, in March of 2004. According to scientists from the National Oceanographic and Atmosphere Administration (NOAA) and Scrips Institution of Oceanography (SIO), the monthly mean CO_2 concentration level peaked at 421 ppm in May 2022.[2] Therefore, this paper will focus on analyze the concentration of CO_2 in March 2004, and compare this change with the change in previous ten years. Then, we will establish three mathematical models, namely ternary linear regression model, Holt-winters model, and simple linear regression model, based on the data from the past to first help describe the concentration level of carbon dioxide in the past. Third, we will use those models to predict how the concentration of carbon dioxide might become in 2100. Fourth, according to OECD's prediction report, it predicts that the concentration of carbon dioxide will reach 685 ppm in 2050,[3] so we will utilize our three models to determine whether the concentration level of CO_2 will reach 685 ppm in 2050 or not. If not, then what the exact year is. According to our modeling, the concentration level of carbon dioxide will not reach 685 ppm in 2050. Based on the three figures of our three models, the exact year that the concentration of carbon dioxide reaches 685 ppm is not identical, so we need to discuss about the accuracy of the three models to determine the most accurate model and result. Finally, the paper will show you the comparison among the three models that are already established to find out which one can predict the change in the concentration levels of CO_2 the most accurate and the reason why this model can predict precisely but the others cannot. However, one thing that needs to be highlighted is that, in order to solve the global warming problem and excessive emissions of carbon dioxide, it is important to find out the relationship between the change in concentration of carbon dioxide and temperature. However, our work just concentrates on analyzing the concentration of CO_2 from the past and predict its change in the future. It doesn't contain any model to analyze the relationship between the temperature and CO_2, so in no way in this paper can we see how the change in CO_2 causes the temperature change.

By making these models to predict how the concentration of carbon dioxide will change in the future, people can see its change in a more visualized and clear way. Therefore, the international community can have a better understanding of emissions so that they can promote countries to take more active measures to deal with the over-emission of harmful gases, since breathing too much CO2 results in high levels of CO2 in the blood associated with a decrease in blood pH (increased acidity) resulting in a condition known as acidosis, a dangerous phenomenon for human, and the health risks will continue to escalate with progressively higher CO2 concentrations. [4] This shows the significance of our modeling.

In general, establishing mathematical models to monitor the global emission of carbon dioxide will bring a far-reaching significance for promoting global response to climate change, promoting sustainable development, and promoting environmental scientific and technological innovation.

2 Main Body

In this section, we would first introduce the formulas and mathematical methods of establishing models to solve the questions mentioned in the introduction, then valuate and provide our discussions about these prediction models.

2.1 Formulas and Methods

The first model we used is the Holt-winter model. The basic principle of the model is the Exponential Smoothing method. It calculates the data in the next time period according to the real data in this time period.

Holt Winter additive model exponential smoothing the seasonal additive model with the seasonal addition method is suitable for predicting time series with the amplitude (height) of the seasonal pattern independent of the average level, or data size is constant.

Additive models are used when there is no trend or sign that the seasonal pattern is dependent on data size. The equations used in the additive model are as follows:

$$S_t = \alpha(X_t - I_{t-L}) + (1-\alpha)(S_{t-1} + b_{t-1}) \tag{1}$$

with:

S_t = Exponential smoothing in year t
S_{t-1} = Exponential smoothing in year $t-1$
b_t = Smoothing trend elements in year t
b_{t-1} = Element smoothing trend on year $t-1$
X_t = Data t
α = Exponential smoothing parameter for data ($0 < \alpha < 1$)
I_{t-L} = Seasonal factor smoothing
L = Seasonal length ($L = 3$, $L = 4$, $L = 6$ or $L = 12$) [5]

According to actual experience, the value of α is generally in the range of 0.1 to 0.3. The larger the value of α, the faster the attenuation rate of the weighted coefficient series, so in fact, the value of a plays a role in controlling the number of historical data participating in the average. a higher value of α means less data is used. Therefore, some basic criteria for choosing value of α can be obtained:

1) If the basic trend of the series is relatively stable, and the prediction bias is caused by random factors, the value of α should be smaller to reduce the correction amplitude, so that the prediction model can contain more information about the historical data.

2) If the basic trend of the forecast target has changed systematically, the value of α should be achieved more. In this way, the original model can be greatly modified to adapt the prediction model to the new change of the prediction target.

The second model we used is the linear regression forecast model. If the actual observed data $(x_i, y_i)(i = 1,2, \ldots \ldots, n)$ is known to be roughly a straight line, then the relationship between the variables y and x can be considered roughly as an approximate linear relationship. In general, these points are not exactly on a straight line, which indicates that the relationship between y and x is not so exact that a given x uniquely determines y. To determine a single linear regression model, the first step is to determine the regression coefficients β_0 and β_1. The following is a least square method to estimate the values of the parameters β_0 and β_1, that is, to determine a set of estimates of β_0 and β_1 such that the regression model is "close" to the linear equation $y = \beta_0 + \beta_1 x$ at all data points $(x_i, y_i)(i = 1,2, \ldots \ldots, n)$.

In order to characterize this degree of "proximity", as long as the sum of squares of the deviation of the observed value of y from the estimate is minimized, that is, only the function is required:

$$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2}$$

This method is called the least square method.

The least square method is used to find the best fit curve or line for one set of data points by reducing the amount of the squares of the offsets (residual part) of the points of the curve. The least square method in the linear regression model use to find b_0 and b_1 predictions such that the cumulative squared distance from the real y_i response $\hat{y} = \beta_0 + \beta_1 x_i$ approaches the minimum of all possible regression coefficients β_0 and β_1 option. [6]

$$(b_0, b_1) = arb \underset{(\beta_0, \beta_1)}{min} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{3}$$

The third model we use is the ternary regression forecast model, also called the cubic curve fitting model. In this model, the main method is called the linear least square method. With this method, we're able to solve the curve fitting problem.

Given a set of two-dimensional data, that is, n points $(x_i, y_i), i = 1,2, \ldots \ldots, n$ on the plane, x_i are not identical, a curve function $y = f(x)$ is sought such that $f(x)$ is closest to all data points under some criterion, that is, the curve fits best.

To solve such a fitting problem, the linear least square method would be effective. We denote $f(x)$ as $a_1 r_1(x) + a_2 r_2(x) + \cdots + a_m r_m(x)$, where $r_k(x)$ is a pre-selected set of linearly independent functions, and a_k is an undetermined coefficient $(k = 1,2, \ldots, m, m < n)$. One thing to note is the least squares criterion: to minimize the sum of squares of the distance δ_i from y_i $(i = 1,2, \ldots, n)$ to x.

Then we need to determine the value of a_k. The procedure is as follows:

$$J(a_1, a_2, \ldots, a_m) = \sum_{j=1}^{n} \delta_j = \sum_{j=1}^{n} (f(x_i) - y_i)^2 \tag{4}$$

$$A = [a_1, \ldots, a_m]^T, Y = (y_1, \ldots, y_n)^T, R = \begin{bmatrix} r_1(x_1) & \cdots & r_m(x_1) \\ \vdots & \vdots & \vdots \\ r_1(x_n) & \cdots & r_m(x_n) \end{bmatrix}_{n \times m} \tag{5}$$

$$A = (R^T R)^{-1} R^T Y \tag{6}$$

When the value of a_k is determined, we just need to get expression formula of the function $r_k(x)$. To select the function $r_k(x)$, there are two situations:

1. If you know a function of y and x, we can determine $r_k(x)$ directly;
2. When the function of y and x cannot be determined, the intuitive choices are:

1) Straight line: $y = a_1x + a_2$
2) Polynomial line: $y = a_1x^m + \cdots + a_mx + a_{m+1}$
3) Hyperbola: $y = a_1/x + a_2$
4) Exponential function: $y = a_1e^{a_2x}$

2.2 Results and Discussions

After considering and analyzing these three questions, we decided to use python code to draw images of the change of carbon dioxide concentration year by year. Through these images, we can accurately obtain the concentration value of carbon dioxide in a specific year and predict the change trend of carbon dioxide in the future. When we show all the figures, we would discuss and evaluate these models.

In the first question, it mainly requires us to find out the change in CO_2 in March 2004, and compare this change with that in previous ten years. Among all the mathematical models, we choose Holt-winter model to visualize the chronological change of carbon dioxide. The figure drew by python is as follows:

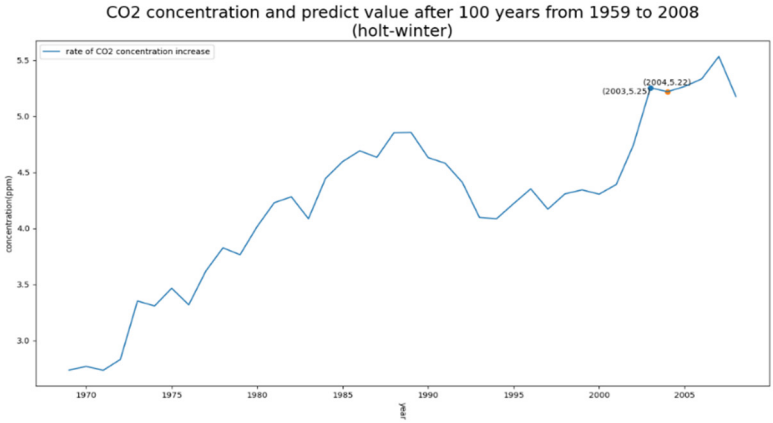

Fig. 1. CO_2 concentration and predicted value after 100 years from 1959 to 2008 by Holt-winter model

As figure 1 shows, the increase rate of CO_2 concentration in March 2004 is 5.22%. While comparing to the increasing rate in the previous 10-year period, all values are less than the rate in 2004 except 2003. The increasing rate of the 2003 actually reaches to 5.25%.

Obviously, the increasing rate in 2003 was higher than that in 2004. Therefore, we disagree with the statement that the March 2004 increase of CO2 resulted in a larger increase than observed over any previous 10-year period.

In the second question, we need to fit three mathematical models to the data to describe past, and predict future, concentration levels of CO_2 in the atmosphere. The three models we choose are Holt-winter model, linear regression model and ternary regression model.

The first model is the Holt-winter model. The figure drew by python is as follows:

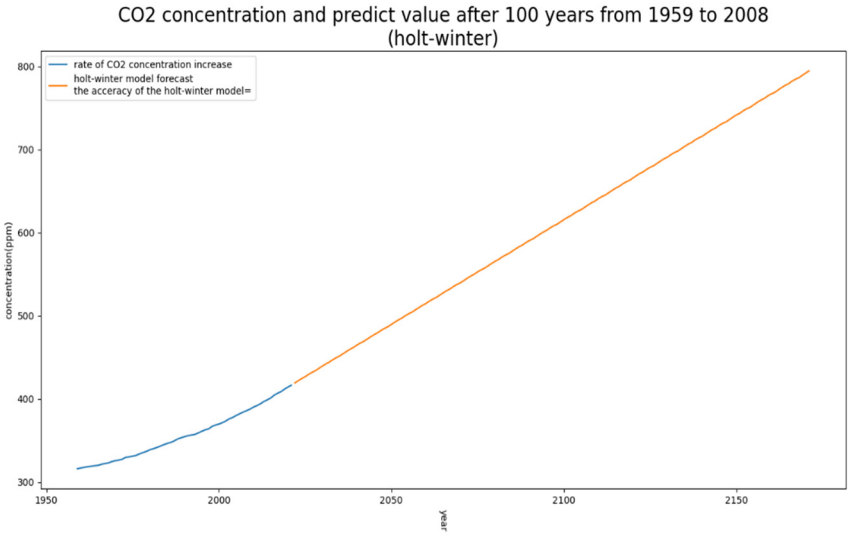

Fig. 2. CO_2 concentration values simulation from 1959 to 2008 by Holt-winter model

As the figure 2 shows, our code first gets the data from 1959 to 2008 and demonstrate it on graph, and we calculate the rate of increase in CO_2 concentration in every year, then we used Holt-winter model to forecast the data after 2008 according to the previous data.

The second model is the linear regression forecast model. The figure is as follows:

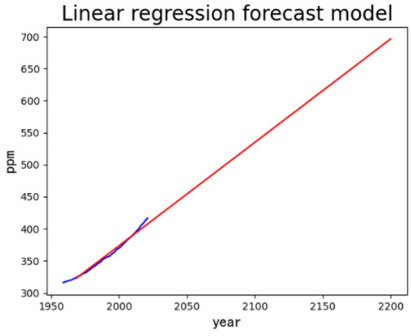

Fig. 3. CO_2 concentration values simulation from 1959 to 2008 by linear regression forecast model

As figure 3 shows, the code calculates the mean value of year and the increase rate, then we get a linear equation from the previous data, and by using the linear equation to forecast the increase rate after year 2008.

The last model is the ternary regression forecast model. The figure drew with this model is as follows:

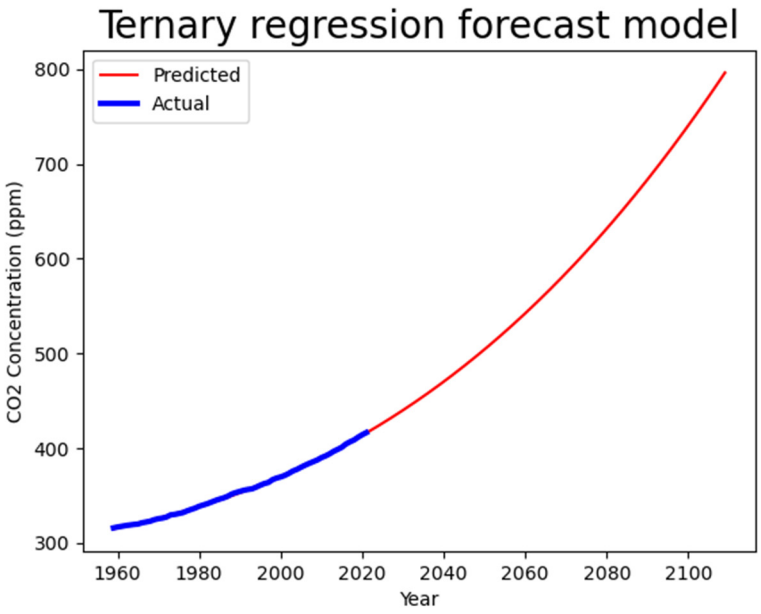

Fig. 4. CO_2 concentration values simulation from 1959 to 2008 by ternary regression forecast model

As figure 4 shows, this model is similar to linear regression, but we get a cubic equation in the end.

As figures 2, 3, and 4 show, the concentration of the carbon dioxide is increasing year by year. Since 1960, the concentration of carbon dioxide has been rising year by year. According to our projections, the concentration of carbon dioxide would continue to increase.

In the third question, we're supposed to use each of our models to predict the CO_2 concentrations in the atmosphere in the year 2100. The three figures drew are as follows:

Holt-winter model:

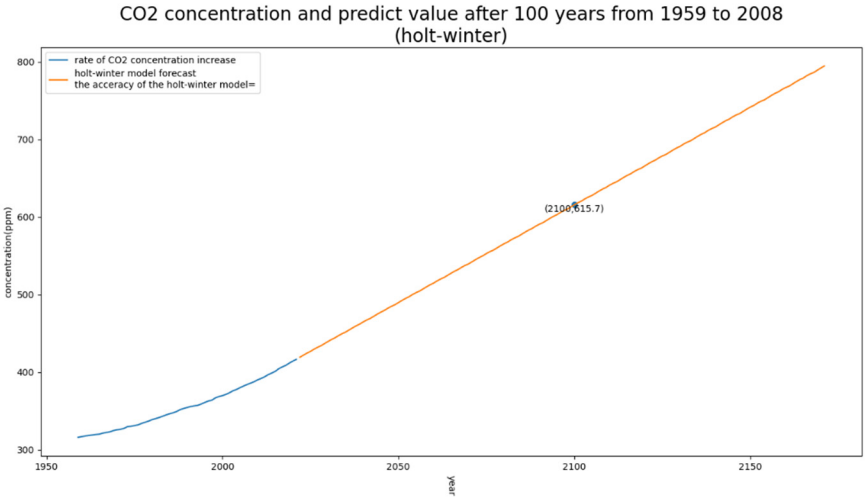

Fig. 5. CO_2 concentration predicted value in 2100 by Holt-winter model

Linear regression forecast model:

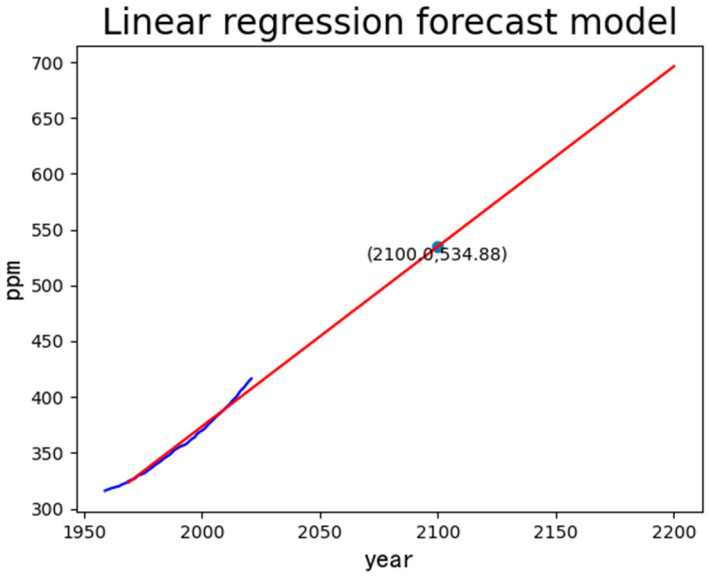

Fig. 6. CO_2 concentration predicted value in 2100 by linear regression forecast model

Ternary regression forecast model:

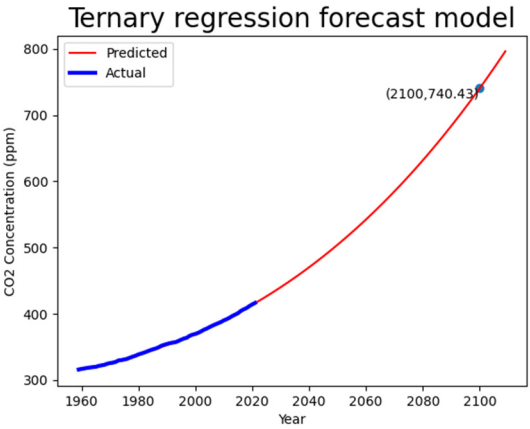

Fig. 7. CO_2 concentration predicted value in 2100 by ternary regression forecast model

As figure 5 shows, the predicted value of Holt-winter model is 615.7 ppm. As figure 6 shows, the predicted value of linear regression forecast model is 534.88 ppm. As figure 7 shows, the predicted value of ternary regression forecast model is 740.43 ppm.

In the fourth question, we need to determine whether the concentration of carbon dioxide would reach 685 ppm in 2050 or not, and if not, then find out the exact year when the concentration of carbon dioxide reaches 685 ppm. The three figures drew by python with each of our models are as follows:

Holt-winter:

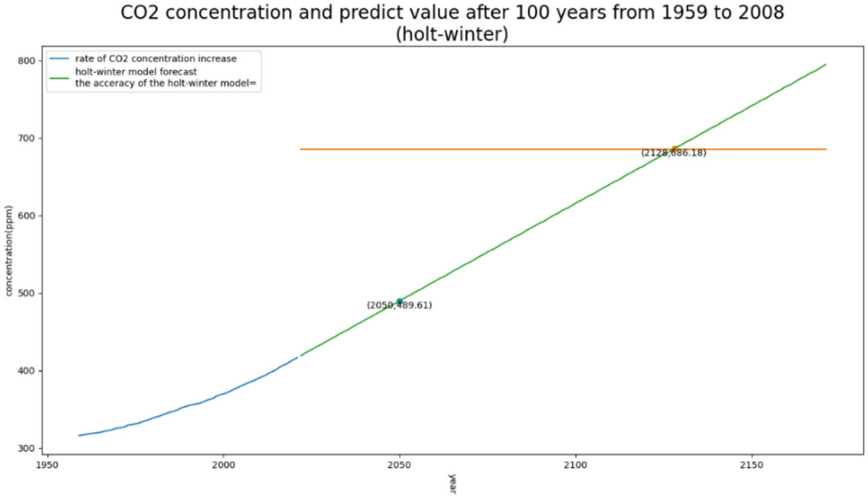

Fig. 8. CO_2 concentration predicted value in 2050 by Holt-winter model

Linear regression forecast model:

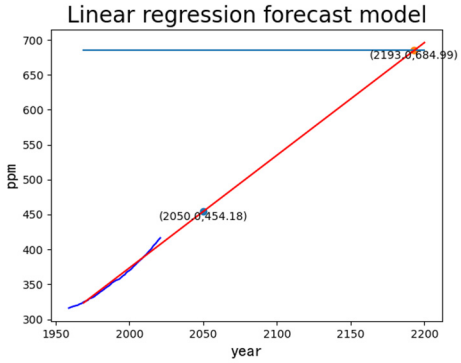

Fig. 9. CO_2 concentration predicted value in 2050 by linear regression forecast model

Ternary regression forecast model:

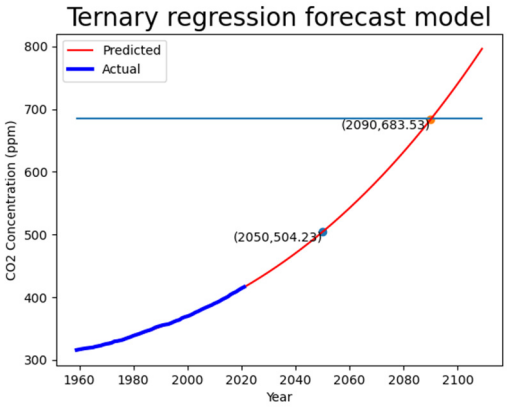

Fig. 10. CO_2 concentration predicted value in 2050 by ternary regression forecast model

As figures 8, 9 and 10 show, the concentration of carbon dioxide doesn't reach 685 ppm in 2050. Instead, as figure 8 shows, it reaches 685 ppm in about 2128; as figure 9 shows, it reaches 685 ppm in about 2193; as figure 10 shows, it reaches 685 ppm in about 2090.

Of course, the reason why these models predict different results is that they have some inaccuracies and errors that need to be discussed and evaluated. Therefore, next, we will interpret and discuss these models.

To evaluate these models, we will apply the coefficient of determination, which is denoted as R^2. The expression of definition of R^2 is as follows:

$$R^2 = \frac{SSV}{SST} = 1 - \frac{SSR}{SST} \tag{7}$$

$$SSV = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = L_{yy}, SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

with:

SSV is the sum of squared variation that can be explained by fitting a straight line;

SST is the sum of squared total variation in the original data y_i;

SSR is the sum of squared residual.

From the definition of the determination coefficient, R^2 has the following simple properties:

1) $0 \leq R^2 \leq 1$;

2) When $R^2 = 1$, $SSV = SST$, that is, at this time, the total variation of the original data can be completely explained by the variation of the fitted value, and the residual is zero ($SSR = 0$), that is, the fitting point matches the original data perfectly.

3) When $R^2 = 0$, the regression equation cannot explain the total variation of the original data at all, and the variation of y is completely caused by factors unrelated to x, in which case $SSR = SST$.

The decision coefficient is an interesting index. On the one hand, it can point out the percentage of explainable variation in the total variation from the perspective of data variation, so as to illustrate the excellent degree of regression line fitting. On the other hand, it can also explain the degree of correlation between the dependent variable y and the fitting variable $\hat{y}$ from the perspective of correlation. From this perspective, the greater the degree of correlation between the fitting variable $\hat{y}$ and the dependent variable y, the higher the goodness of the fitting line.

Therefore, the closer R^2 approaches 1, the more accurate and reliable the model is. Here are the figures that show the value of R^2 of each of our models:

Holt-winter model:

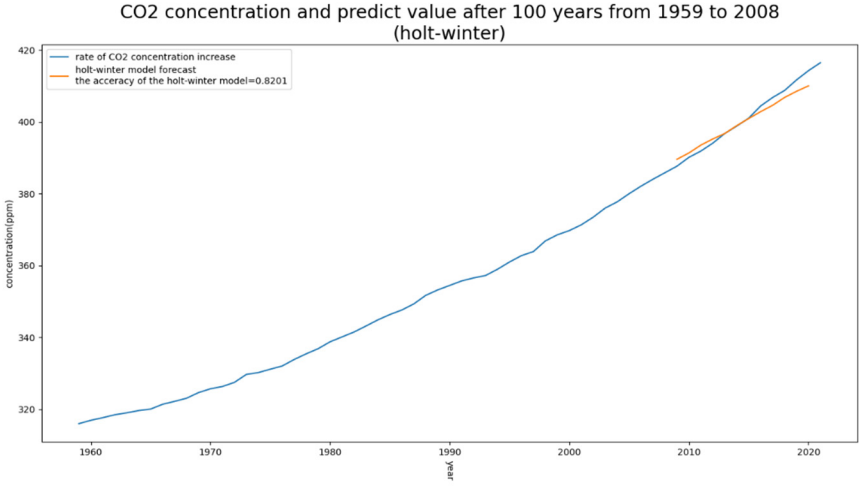

Fig. 11. Value of R^2 of Holt-winter model

Linear regression forecast model:

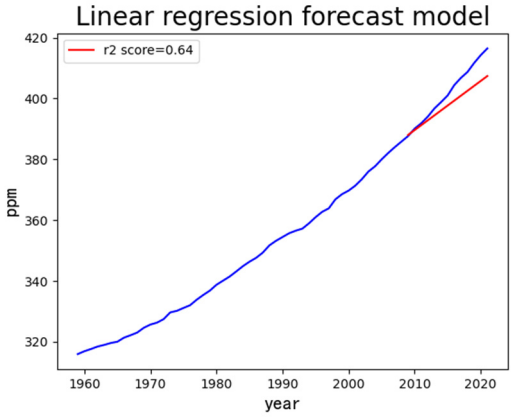

Fig. 12. Value of $\boldsymbol{R^2}$ of linear regression forecast model

Ternary regression forecast model:

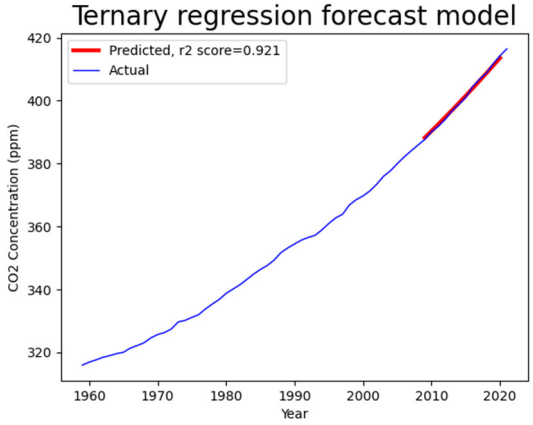

Fig. 13. Value of $\boldsymbol{R^2}$ of ternary regression forecast model

As figures 11, 12 and 13 show, the value of R^2 that most approaches to 1 is the R^2 value of the ternary regression, which reaches to 0.921. As a result, the most accurate model is the ternary regression forecast model.

3 Conclusion

Based on the models, it is obviously to see that the change in concentration level of carbon dioxide is not the largest in March 2004 but in 2003. Then, by establishing three models using data from previous years, we notice that the concentration of CO_2 increases rapidly in the future.

1) The concentration level of CO_2 reaches 618.5ppm according to Holt-winter model in 2100

2) It reaches 534.88ppm in 2100 according to linear regression forecast model

3) It reaches 740.43ppm in 2100 according to ternary regression forecast model

At last, we determine that the concentration of carbon dioxide doesn't reach 685 ppm in 2050 according to three models.

1) Its concentration reaches approximately 685ppm in 2128 according to Holt-winter model.

2) Its concentration reaches approximately 685ppm in 2193 according to linear regression model.

3) Its concentration reaches approximately 685ppm in 2090 according to ternary regression model.

After all, the ternary regression forecast model is the most accurate model that was used in this work.

All of the result shows the global warming will become intensive in the future without any human intervention. Therefore, this paper can not only make readers be aware of environmental problems, but also provide governments and environmental protection organizations data to deal with the global warming.

Finally, it is also an important thing to point out that the mathematical models that are used in this paper can also be applied in other areas of studies that need to do a prediction, not just predicting concentration of CO_2. For example, it can be used to predict the future economy market trends, the number of wild animals, the degree of resource consumption. However, since these models are built based on some assumptions that we can't guarantee they won't be broken in the future, there will always be some error and uncertainty.

References

1. Fanhe Kong, Guanhe Rim, MinGyu Song, Cornelia Rosu, Pranjali Priyadarshini, Ryan P. Lively, Matthew J. Realff, Christopher W. Jones, 2022. Research Needs Targeting Direct Air Capture of Carbon Dioxide: Material & Process Performance Characteristics Under Realistic Environmental Conditions. https://www.osti.gov/servlets/purl/1970455
2. Siqi Yang, Haoran Yao, Yuxuan Sun, HanFeng Cai, 2023. Keep an eye on global temperature changes. https://madison-proceedings.com/index.php/aetr/article/view/1107
3. Green T. 2013. Trends Shaping Education 2014 Spotlight 4. document (psu.edu)
4. P.N. Bierwirth, 2019. Carbon dioxide toxicity and climate change: a major unapprehended risk for human health. https://www.sepanso33.org/IMG/pdf/co2toxicity.pdf
5. Dastan Maulud, Adnan M. Abdulazeez, 2020. A Review on Linear Regression Comprehensive in Machine Learning. https://jastt.org/index.php/jasttpath/article/view/57
6. Nurhamidah N, Nusyirwan N, Faisol A, 2020. Forecasting seasonal time series data using the holt-winters exponential smoothing method of additive models. Forecasting Seasonal Time Series Data using The Holt-Winters Exponential Smoothing Method of Additive Models.pdf (unila.ac.id)