Forecasting Future Values of the Atmospheric Concentration Levels of CO2

Coleton Reitan

12/16/2021

QAC320 Final Paper

Introduction

This paper was designed to emphasize the rate at which the globe has been experiencing an increase in atmospheric levels of carbon dioxide as well as display values for which these levels may rise to in future years.

Increasing global atmospheric levels of carbon dioxide (CO2) can be dated back to the industrial revolution. Although this was a time when many important technological advancements were made, within those advancements was the introduction of the burning of fossil fuels, primarily coal and oil, which are emitters of CO2. Since then, CO2 emission levels have been steadily increasing, resulting in an increase in the concentration of CO2 within the Earth's atmosphere. Even as CO2 levels continue to rise, the primary concern is not the gas itself, but the consequences the gas trapped within the atmosphere has on the planet. These consequences have become a term known today as climate change.

To forecast the rate at which atmospheric CO2 levels are increasing globally, this paper will be performing a time series analysis on the atmospheric concentration of CO2 based on data from an observatory located in Mauna Loa, Hawaii. The use of simple forecasting models such as Holt-Winters seasonal smoothing and moderately advanced forecasting models such as an autoregressive integrated moving average (ARIMA) model will be used to determine these forecasted values. By the end of this analysis, there should be an understanding of not only the rate of increase of which CO2 levels are rising globally, but also where these levels could potentially rise to in coming years. By the end of this analysis, there should be an understanding of not only the rate of increase of which CO2 levels are rising globally, but also where these levels could potentially rise to over the next several years.

1.1 Effects of Climate Change on Weather

Climate change has quickly become an international topic of concern. Over the past decade, there has been a substantial increase in global "extreme weather events", which are defined by weather events unsimilar to 90-95% of previous weather events in a certain area (national academies, 2019). Across the globe, numerous studies have shown a link between extreme weather events and climate change, such as extreme drought in the American Southwest, record high temperatures in Western Europe, and Winter precipitation in Beijing (BAMS, 2020). After many analyses of extreme weather events and their link to climate change, it is evident that climate change is quickly becoming the world's largest problem and will continue to be a primary topic of concern for years to come.

1.2 Relevance of CO2 in Climate Change

To evaluate the severity of climate change to come, a good indication that researchers can look at is the greenhouse gas1 levels found within the Earth's atmosphere. Although greenhouse gases can be naturally found and are a necessary aspect of life on Earth, a high concentration of greenhouse gases within the atmosphere can result in undesirable outcomes, which in turn results in global warming and climate change.

In a 2019 study conducted by the Center for Climate and Energy Solutions, the United States alone produced over 6.6 billion metric tons of greenhouse gases. Although too much of any greenhouse gas could have negative effects on the planet, CO2 seems to be the primary problem, as emissions for this gas are dangerously high. In the same study, it was found that of the 6.6 billion metric tons of greenhouse gases emitted by the United States, 80% was CO2 emissions. Evidently, there is no shortage of CO2 being emitted and naturally as more of this gas is emitted, more is being trapped within the Earth's atmosphere. Many observatories around the world have proved these emissions are staying in the Earth's atmosphere, as the atmospheric concentration of CO2 has been shown to be continuously rising at an undeterred rate.

1.3 Literature Review

Much of the research performed in this field was conducted under the basis of forecasting future CO2 emissions, rather than future atmospheric concentrations of CO2. Although the reasoning behind this is because through the determination of future values for CO2 emissions, one can derive an estimate of future atmospheric concentration levels of CO2, these calculations may not be as accurate as forecasting the atmospheric concentration levels of CO2 directly from historical data. This paper aims to add to the literature by reporting forecasted model results based on historical data of atmospheric CO2 levels, rather than an estimation based on emission levels. However, for the purpose of this paper, the modeling methods used throughout this previous research remains valuable.

In a study which forecasts future values of CO2 emissions in Bangladesh (Hossain et al., 2017), the researchers performed a time series analysis using modeling methods such as Holt-Winters non-seasonal smoothing, artificial neural networks, as well as an ARIMA model. By comparing the models with specific model feedback tests such as Akaike Information Criterion (AIC) and root mean square error (RMSE), it was determined that the ARIMA model displayed the most accurate forecasted values. Such information will be valuable to know, as this paper displays the use of the modeling methods of Holt-Winters smoothing and ARIMA, however seasonal aspects will be added to both models. Another study which forecasts future CO2 emissions in Iran (Lotfalipour et al., 2013) also utilizes an ARIMA model to forecast future CO2 emissions, as well as Grey System modeling. Using Grey System modeling, the researchers were able to forecast future values of CO2 emissions by focusing on the association between model structure and certain conditions. These conditions were expanded on to include uncertainty, multi-data input, discrete data and lack of data. Even as this modeling method will not be used in this paper, the method of comparing models through the RMSE will be, as this is how the researchers determined that the Grey System modeling showed more accurate results than the ARIMA modeling.

Since the data being used in this research was collected in Mauna Loa, Hawaii, it is important to understand if there is a difference of data collection of atmospheric concentration levels of CO2 between outland observatory sites (such as Mauna Loa), and inland observatory sites. In a study conducted to determine the difference in the data collection of atmospheric CO2 levels based on spatial variation across 7 observatories (inland and outland) in China (Zhang et al., 2008), data collection techniques found that the data was significantly similar. Certain characteristics of atmospheric CO2 levels such as seasonality and trend were found to be constant across all data collection sites, as well as the CO2 values themselves being statistically similar. This will be important to this paper, as an outland site such as Mauna Loa can be seen as a collection site for a very large area mass.

1.4 Modeling Techniques

To have a basic understanding of possible future values of the atmospheric concentration of CO2, simple forecasting methods were applied and used as benchmarks for further modeling techniques such as SARIMA and SARIMAX models. Three simple forecasting methods were applied: the naïve method, the drift method, and Holt-Winters smoothing method. Each method has its own strengths and weaknesses, displaying a solid range of forecasting methods, giving preliminary readings as to possible intervals in which further modeling techniques can be tested against. As mentioned, the two moderately advanced methods of forecasting being used in this paper will be seasonal ARIMA (SARIMA) and seasonal ARIMAX (SARIMAX) models.

1.4.1 Simple Forecasting Techniques

Through the use of the naïve method, forecasted values will be equal to the value of the last observation. However, since there is evidence of seasonality within the dataset, there will be a correction for seasonality. This is done by setting the forecast equal to the last observed value from the same season of the year (for example, the same month of the previous year). The drift method is a variation of the naïve method, with the difference being the in the method's ability to forecast an increase and decrease (change) over time. This is important because where the naïve method may lack trend, the drift method makes up for. However, a major downfall for the drift method is its inability to effectively account for seasonality. The Holt-Winters method will be return the most accurate forecast model (of simple methods) because of the three smoothing equations added to it: one for level, one for trend and one for seasonality. In addition, there are two versions of this method which, when applied correctly, can have significant effects: the additive and multiplicative versions. Since there is an apparent additive trend to the data, the additive version of this method will be applied.

1.4.2 Autoregressive Integrated Moving Average Models

Beyond simple forecasting methods, this paper will be evaluating the data to estimate parameters that can be used to create a reliable ARIMA model, more specifically, due to the evident seasonality in the data, a SARIMA model. However, before estimating model parameters, the data will first be checked and fitted for stationarity. ARIMA forecasting models stand out from simple forecasting models for several reasons, each of which encompass an idea from the simple forecasting methods. The model uses lagged moving averages which in turn smooths the data, the model assumes that previous datapoints will resemble future datapoints and, most importantly, the model parameters can be specified by the researcher themselves. The advancement from an ARIMA model is to look towards an ARIMAX model. An ARIMAX model extends from the ARIMA model, but also uses autocorrelation in residuals with an exogenous variable which in turn can create a more accurate forecasting model.

Methods

Assessing forecasted values of the atmospheric levels of carbon dioxide in the Earth's atmosphere will continually prove to be an important aspect of policy and decision making by national leaders in years to come. However, it is imperative that these forecasted values are as accurate as possible and do not underestimate nor overestimate possible outcomes, as this could not only lead to improper decision making, but also failure in efficiency of preparation for said decisions. In this study, R program software will be used to assess the applied methods.

To precisely assess the forecasted models, there must be an understanding of what the data consists of as well as how it was collected. For the data being used in this study, it is important to note that atmospheric carbon dioxide levels are measured in parts per million and has undergone several on site tests and evaluations to produce the given values (see section 2.1). Carbon dioxide levels across different sites may produce different values, but as discussed through the research of Zhang et al, values found show similar characteristics and significant similarity. Although there are significant similarities across data collection sites, national decision and policy makers should not only make assessments based on global research, but local research as well.

In forecasting future values with this research's applied methods, it is important to note that data transformation must be conducted to fit a time series analysis (see section 2.2). It should also be noted that in the data transformation, the data has been broken down into a training and testing groups to produce more accurate forecasted values. After the data has been transformed, data exploration such as decomposition to determine characteristics of the data is performed. Said data exploration allows certain parameter modeling to be more accurate, which also leads to a more accurate model.

2.1 Data

This paper utilizes data from the Global Monitoring Laboratory, located in Mauna Loa, Hawaii and was collected by Dr. Pieter Tans and Dr. Ralph Keeling. The data is composed of monthly averages for atmospheric levels of CO2 measured in parts per million (PPM), and dates as far back as 1960. Mauna Loa serves as a reliable data collection site for two reasons: (i) it's height above sea level is about 3,400 meters, allowing the readings to be taken directly from the upper atmosphere and enables the readings to be more accurate. (ii) Mauna Loa is very distant from any major pollutants, so the atmospheric quality levels are true on a planetary scale. Aside from the geographic reliability of this site, the site also uses state of the art technology developed to give very precise and accurate measurements, along with constant hands on calibrations being made to ensure the technology is always fully functioning.

2.2 Data Transformation

In transforming the data, it is important to note the data was put into a time series format through the use of the ts() function in R. Once done, the data was then broken into two groups; the training group and the testing group. Since the data goes back to 1960, the training group encompassed data from 1960 to 2015, whereas the testing group used data from 2015 to 2021. It is important to note that about 80% of the data should be used as a testing group, as it will allow the forecasted values to be as accurate as possible.

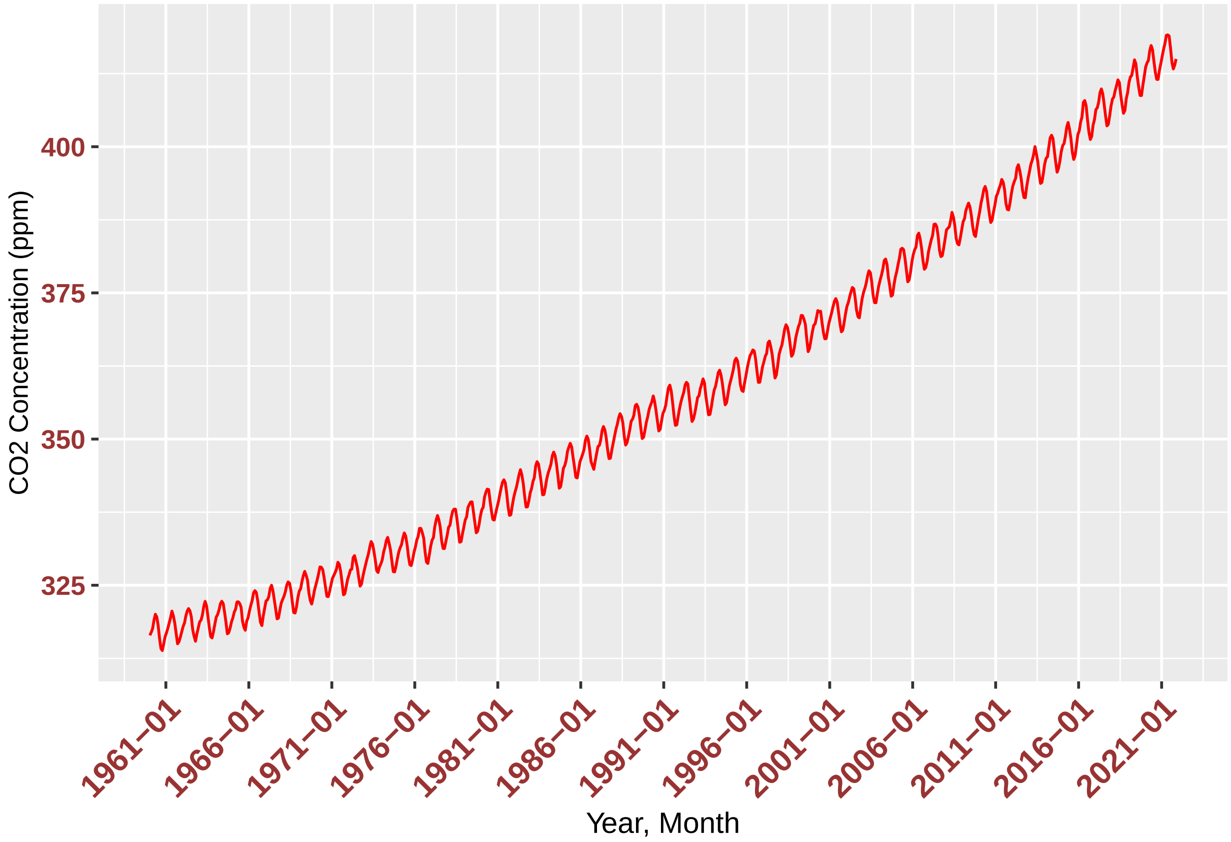

Figure 1: Plot of Atmospheric CO2 Levels Since 1960

As seen in Figure 1, atmospheric CO2 levels since 1960 have shown to have a consistent positive additive seasonal trend.

2.3 Simple Forecasting Models

The use of simple forecasting models such as the naïve, drift and mean methods (see section 1.4.1) allowed the ARIMA models (see section 1.4.2) to have benchmark models to follow. Each simple forecasting method has its own equation used, with slight adjustments made to better suit the data. Each method function is from the R package forecast, with the corresponding call functions shown.

The naive model equation is $\mathrm{Y}_{\mathrm{t+h|T}}=Y_{\mathrm{T+h\text{-}m(k+1)}}$ R function *snaive*

The drift model equation is $\mathrm{Y}_{\mathrm{t+h|T}}=Y_{\mathrm{T}}+(\mathrm{h/T\text{-}1})\sum_{t=2}^{T}(Y_{\mathrm{t}}\text{-}Y_{\mathrm{t\text{-}1}})=Y_{\mathrm{T}}\mathrm{h}(Y_{\mathrm{T}}\text{-\_Y\_1/T-1})$ R function *rwf*

The Holt-Winters model equation is $Y_{\mathrm{t+h|T}}=\iota_{\mathrm{t}}+\mathrm{hb}_{\mathrm{t}}+\mathrm{s\text{~}t+h\text{-}m(k+1)}$ R function *holtwinters*

2.4 ARIMA Modeling and Parameter Selection

Before estimating model parameters, the data must first be checked and fitted for stationarity. This was done by using an augmented Dicker-Fully test, which showed results of non-stationary data. To correct for this, a first order difference was applied to the data. After reattempting the ADF test with the differenced data, the data showed stationary characteristics. Further evaluation of data characteristics lead to the creation of the SARIMA model ARIMA(1,1,3)(1,1,2).

With the equation being $(1\text{-}\phi 1B)(1\text{-}\phi 1\ \mathrm{B12})(1\text{-}\mathrm{B})(1\text{-}\mathrm{B12})\mathrm{Y}_{\mathrm{t}}=(1+\theta\ 1\mathrm{B})(1+\Theta\ 1\mathrm{B12})\epsilon\ \mathrm{t}$

For the SARIMAX model, the exogenous variables being used was average income in the United States as well as United States unemployment rate. These variables may provide some feedback that may be interesting to examine. The equation of the SARIMAX model is very similar to that of the SARIMA model, the only difference is the addition of an exogenous covariate.

With the equation being $(1\text{-}\phi 1B)(1\text{-}\phi 1\ \mathrm{B12})(1\text{-}\mathrm{B})(1\text{-}\mathrm{B12})\mathrm{Y}_{\mathrm{t}}=(1+\theta\ 1\mathrm{B})(1+\Theta\ 1\mathrm{B12})\epsilon\mathrm{t}+\beta\ \mathrm{xt}$

2.5 Model Reliability

After every model returned forecasted values, each model had its residuals checked in an effort to determine whether the model had done an appropriate job of capturing information within the data. Through the R software function *checkresiduals()*, an Ljung-Box Test was performed and displayed important statistical information as to whether the model adequately displayed the captured information in the data. Models that displayed a p-value greater than or equal to .05 from the Ljung-Box Test were deemed statistically appropriate, while those with a p-value less than .05 were deemed unreliable.

Results

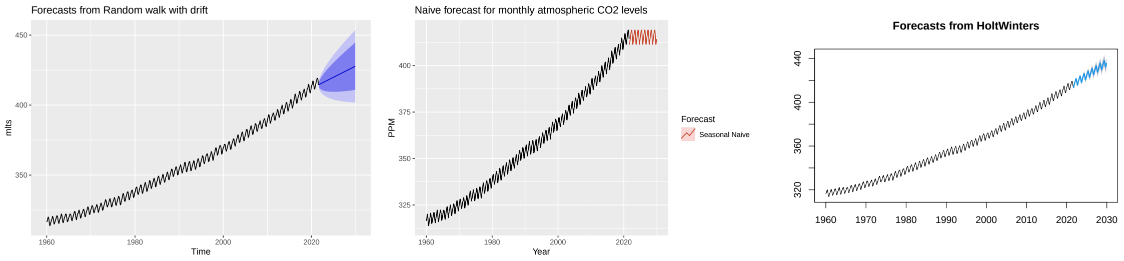

Figure 2: Simple Forecasting Methods Plots

3.1 Simple Forecasting Output

The simple forecasting model outputs showed varied results, but each proving to be a good benchmark for the ARIMA model outputs (see section 3.2). The naïve method model did not maintain the trend, but did keep seasonality, as can be seen in the plot above. The drift method model did a good job of projecting confidence intervals in which possible forecasted values could be seen but failed to keep seasonality. The Holt-Winters method model proved to be the best model of the three, as seasonality and trend were kept and by examining the plot, stayed in line with the moving average.

3.2 ARIMA Forecasting Output and Comparisons

Recall the model that was decided to be used for the manual SARIMA model was an ARIMA(0,1,3)(3,1,1)[12]. Through examination of the plot of the SARIMA model, it is evident that the SARIMA model did a good job of keeping trend as well as seasonality, while predicting values up to the year 2030. Looking at the Akaike information criterion (AIC) the model gave a value of -1095.1323, which compared to other models that it was tested against is relatively good.

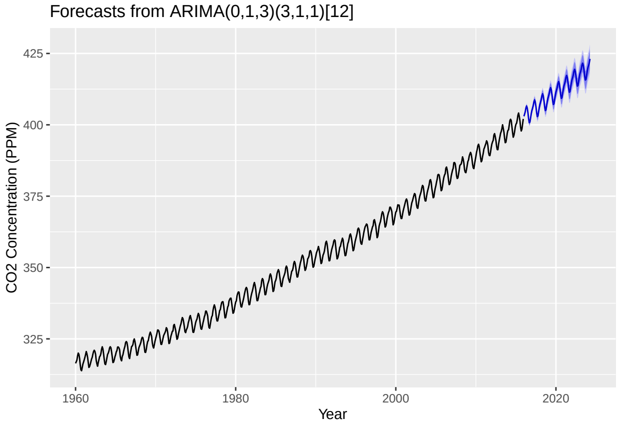

Figure 3: Manual SARIMA Plot

Comparing the model that was created by manual parameter selection to an autoARIMA model may prove to be beneficial, as the R software may be able to identify certain aspects of parameter specification that was misunderstood. The model parameters used by the autoARIMA function was ARIMA(1,1,1)(0,1,1)[12], which produced an AIC of -1101.2607. Although the AIC of the autoARIMA model was better than the manual parameter selection model, the discrepancy is too small to deem significant. However, the autoARIMA model was still proved to be the more accurate of the two models.

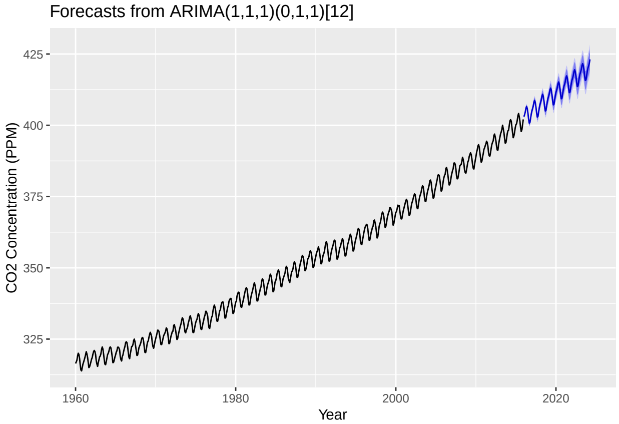

Figure 4: Auto SARIMA Plot

Among the two models described, there were other SARIMA models which were used to compare, all of which have parameters that were manually adjusted to find a more accurate model.

```
##                Model_name       AIC
## 1 ARIMA(0,1,3)(3,1,1)[12] -690.5971
## 2 ARIMA(0,1,1)(3,1,1)[12] -690.9743
## 3 ARIMA(1,1,0)(1,1,0)[12] -501.1924
## 4 ARIMA(1,1,2)(1,1,0)[12] -515.4935
## 5 ARIMA(1,1,3)(0,1,1)[12] -694.6569
## 6 ARIMA(1,1,1)(1,1,0)[12] -517.2885
## 7 ARIMA(1,1,1)(1,1,0)[12] -517.2885
## 8 ARIMA(1,1,0)(1,1,1)[12] -682.3033
## 9 ARIMA(1,1,1)(0,1,1)[12] -697.5774
```

Of the 7 other SARIMA models used to find a better model specification, only one showed to have a lower AIC than the primary manual model, which was the model ARIMA(1,1,3)(0,1,1)[12], which had an AIC of -1098.8099. Which again has a very small difference from the AIC of the original manual SARIMA model. It is interesting to note that both model selections with a lower AIC than the original manual parameter SARIMA model have p=1, while the original had p=0. This may show that a mistake was made in examining the ACF plot of the stationary data.

3.2.2 ARIMAX Forecasting Output and Comparisons

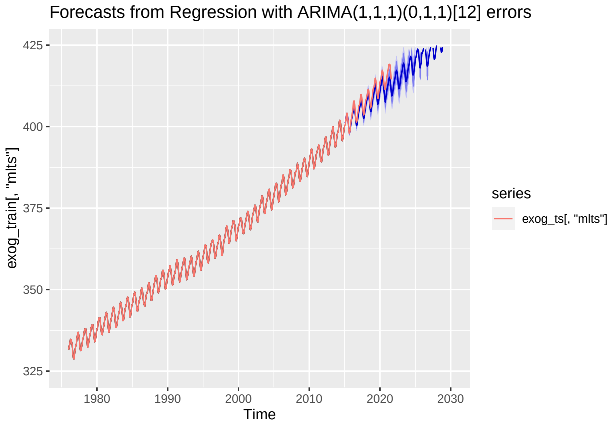

Figure 5: ARIMAX plot with US average income as exogenous variable

The exogenous variable used in examining the first SARIMAX model was average income in the United States. Through plot examination, both the seasonality and trend are kept, showing a strong resemblance to the original data points. In order to obtain the model specifications, the autoARIMA function in R was used and produced a model with parameters ARIMA(0,1,1)(2,1,0)[12]. The resulting AIC from this model was 283.94. For the second tested SARIMAX model, recall that the exogenous variable United States unemployment was tested with the model. It should be immediately pointed out that this model showed a p-value of below .05 in the Ljung-Box Test, which allows the assumption that the model did not do well enough capturing information within the data, showing too much correlation within the random terms.

Discussion and Application

The focus of this paper was to provide a framework for forecasting future values of the atmospheric concentration levels of carbon dioxide using multiple time series analysis techniques in which future works can base their models after.

As discussed in the former part of the paper, previous studies investigated forecasting values of carbon emissions using similar models, but then would obtain estimates of future atmospheric CO2 levels based on forecasted emission levels. Although it is hard to make an assumption of whether there was an agreeance between the estimated values based on forecasted CO2 emissions and forecasted values for atmospheric levels of CO2, there were similarities shown in the plots of the projected values of CO2 emissions and atmospheric levels of CO2. The modeling methods used in this paper showed very similar trajectories of continual increase into the near future, with certainty. With the forecasted values collected, it has been shown that by 2030, atmospheric levels of CO2 will reach up to 420ppm.

This paper added to the literature by displaying an array of forecasting modeling techniques that could be used to determine future atmospheric levels of CO2. Future studies should attempt to display a comparison between estimated atmospheric CO2 levels based on forecasted CO2 emissions and forecasted atmospheric levels of CO2 based on the historical atmospheric levels of CO2. In doing so, this would tell whether CO2 emissions or atmospheric levels of CO2 are a more telling sign of future CO2 levels and possibly create more opportunity to diminishing climate change before it is too late.

Conclusion

Through the methods used in this paper, successful forecasting models were created in which could determine future values of atmospheric levels of carbon dioxide. Unfortunately, the predicted values proved to be continually increasing over the next decade, with each year reaching a new record high for atmospheric levels of CO2.

National policy and decision makers should focus on carbon dioxide (and greenhouse gas) research and make decisions that are heavily influenced by said research. If this type of research continues to be overlooked and put off, the globe will continue to experience extreme weather events that will likely prove to be more extreme and common with time.

Acknowledgments

I would like to give a very grand thank you to Professor Maryam Gooyabadi, for always bringing a positive attitude and cheerfullness to class, even though it was at 8:50 in the morning. This class was truly enjoyable, and the skills I've gained from it will be forever in my possession. Thank you.

References

(national academies, 2019) Global warming is contributing to extreme weather events | National Academies (BAMS, 2020) Explaining Extreme Events from a Climate Perspective - American Meteorological Society (ametsoc.org) (Center for Climate and Energy Solutions) U.S. Emissions | Center for Climate and Energy Solutions (c2es.org) (Hossain et al., 2017) Sample style (ru.ac.bd) (Lotfalipour et al., 2013) download (psu.edu) (Zhang et al. 2008) Temporal and spatial variations of the atmospheric CO2 concentration in China (wiley.com)