Time Series Analysis on Climate Change

India, Colombia, Australia, California, New York, Texas

Eden Belay, Shakeeb Habash, Beza Lemma, Jiajia Liu, Xuefei Qiao, Kexuan Song

Abstract

Climate change has been a controversial topic in many aspects, such as whether climate change is really happening and if so to what extent? Another question often posed is whether climate change is caused by a natural progression or is affected by human activity. With these questions in mind, we found a dataset containing monthly average land temperature at the global, country, state, and city level from 1750 to 2013. We conducted preliminary research and selected three representative countries - India, Colombia, Australia - and three representative states – California, New York, and Texas - as our study objects. Through time series analysis, we discovered that the average land temperature of all countries and states has been increasing over the past 250 years, but at a slow rate. We also compared the accuracy of ETS, ARIMA, neural network, and bagging models and selected the best candidate to forecast the average land temperature for those countries and states. In addition, we brought in a yearly CO_2 and greenhouse gas emission dataset and found out that human activity may be positively correlated with the temperature increase.

Index Terms: Temperature, ETS, ARIMA, neural network, bagging, emission

1 Introduction and Motivation

The purpose of this report is to forecast climate change. Hot days and heatwaves are becoming increasingly common in all geographical areas; 2020 was one of the warmest years on record. Temperature changes can also lead to increases in rainfall. As a result, storms become more severe and frequent and can damage land and human lives. More areas are experiencing water scarcity. Droughts can cause devastation by causing massive sand and dust storms that can move billions of tons of sand across continents. Predicting climate change will help us understand how climate is changing and will also be helpful to prepare for future natural disasters. Through the analysis of this paper, we will compare the trend and seasonality of three countries and three states. Then we will move on to build and compare several forecasting models. Lastly, we will also compare the greenhouse effect as it is related to global warming and climate change.

2 The Data Set

Our global earth surface temperature data set contains monthly data on land average temperature, maximum temperature, minimum temperature, land and ocean average temperature between 1850 and 2015. In addition, we have separate temperature data sets on the country level and city level worldwide.

We will start with exploring global temperature change, then move on to several selected countries and cities. We also consider adding data sets of CO2/Green House Gas emissions, GDP by country in order to discover whether some of these factors have been contributing to climate change.

Furthermore, we'll determine the training and test dataset for all countries with the window () function and do the test on our training datasets to see if they're linear. If the test shows the linearity, we can use the training dataset to build stl, ETS, and ARIMA models to forecast values, otherwise, we'll use neural network to forecast. Lastly, we'll check the assumptions of all models and pick the winner for each country through some metrics on the test dataset, including MAPE, MASE, AIC, etc.

To better understand how the changes in temperature spread globally, instead of analyzing the global temperature as a whole, we intentionally select three coutries to represent different longitude and latitude combinations:

- ❑ India is located in both the Northern and Eastern hemispheres
- ❑ Colombia is located in both the Southern and Western hemispheres
- ❑ Australia is located in both the Southern and Eastern hemispheres

In addition, we would like to further investigate how temperature changes within one country. Thus in this case, we select three states from different parts of the United States:

- ❑ New York State from the Northeast
- ❑ California from the Southwest
- ❑ Texas from the South

3 Exploratory Data Analysis

In this section, we first observe polar plots and trend plots for original time series of three countries and three states in the U.S. from July 1852 to Aug 2013. We detect outliers and missing values and complete the cleaning procedures. Next, we conduct STL decomposition and calculate strength of trend and seasonality to verify if what we saw from previous plots in the beginning is correct. Finally, we use an unsupervised tree to see the similarity on the cleaned time series with both Euclidean and Correlation distance metric.

3.1 Trend and Seasonality Analysis

Before cleaning the dataset, the plot Figure 1. below, it shows that both India and Australia have no trend, and Colombia shows an upward trend, and there are some missing values in the India time series. In the Colombia panel, there is one point - purple highlighted - that we think it may be an outlier because that value is too big and out of the regular range.

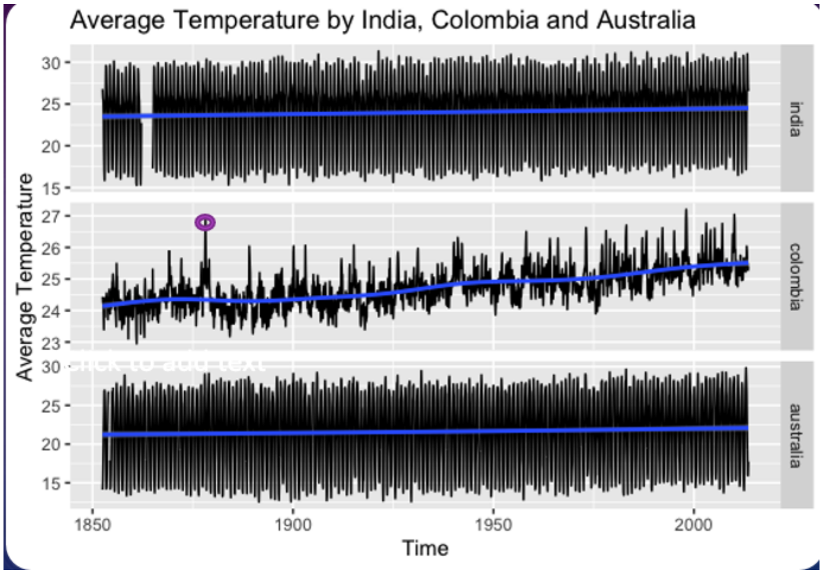

Figure 1. Trend analysis for three countries

From Figure 2 below, we can see that all three states including California, New York and Texas don't show a trend.

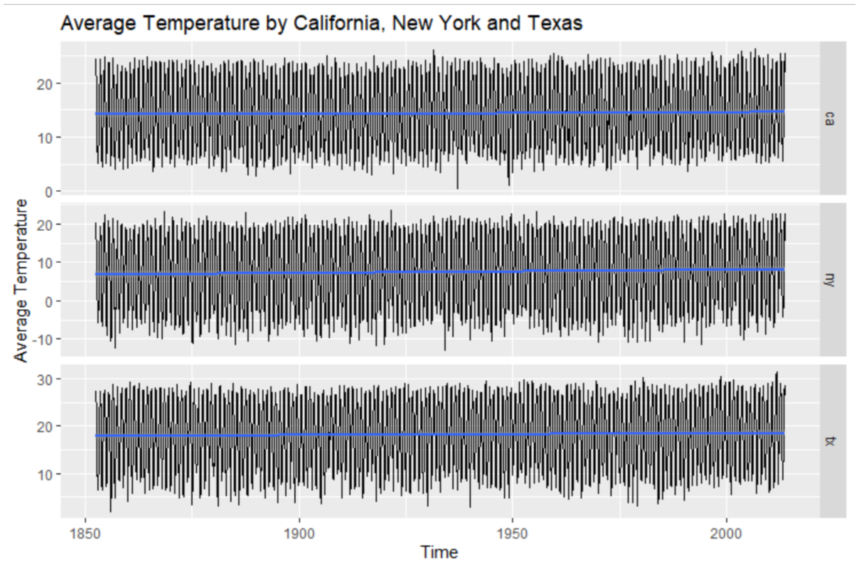

Figure 2. Trend analysis for three states

Furthermore, the polar plots below show that seasonality in India time series is very strong, and the average temperature in May and June is the highest, and in January is the lowest almost every year, and no colorful separation (show the trend) can be found there.

In Figure 3, the polar plot for Colombia, there is a strong colorful separation, but it does show an upward trend in the time series of Colombia (the color is darker in the inner ring and lighter in the outer ring). But its seasonality is not as clear as the previous one.

In the polar plot for Australia, the seasonality is as strong as that in India, and the average temperature in January or February is the highest, and in July is the lowest, which is opposed to the months that has the highest and lowest temperature in India. It may be caused by their different hemisphere locations. No strong colorful separation can be seen in the plot.

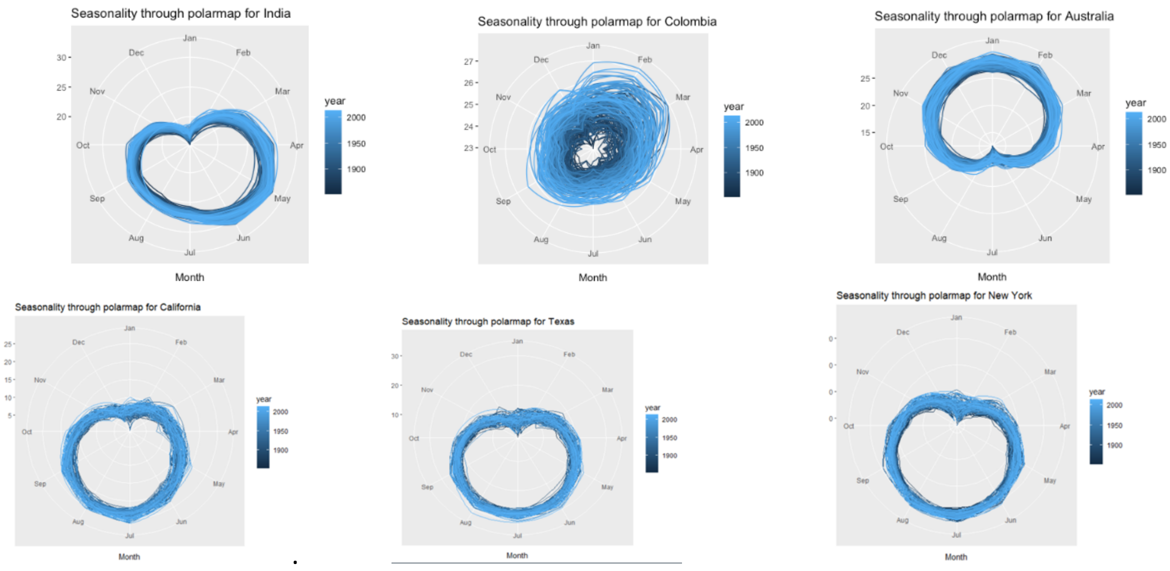

Figure 3. Seasonality Analyses for three countries and three states

For three US states including California, Texas, and New York, all of them show strong seasonality but no colorful separation (trend). January has the lowest average temperature for those states, and July or August has the highest average temperature for almost every year for all three states.

3.2 Dataset Cleaning

Given that our target is the average temperature, and it is reasonable to have both positive and negative values in our dataset, so we did not transform those values with log() method. Through combining above trend and polar plots, we can find that there are some missing values in India time series, and outliers in Colombia time series. Through the objective formula shown below, we can find that there are about 33 missing values in India time series, and more outliers in that series as well.

For the three states in the U.S., we do not have any missing values, but there are six outliers in New York. These outliers are periods in the winter where the temperature has fallen below –10.

After cleaning the dataset with replacing outliers and filling in missing values with lambda=’auto’, we have those plots below as a comparison between cleaned and uncleaned time series for three countries and three states, and in Colombia we can find that the outlier we mentioned before has been replaced here.

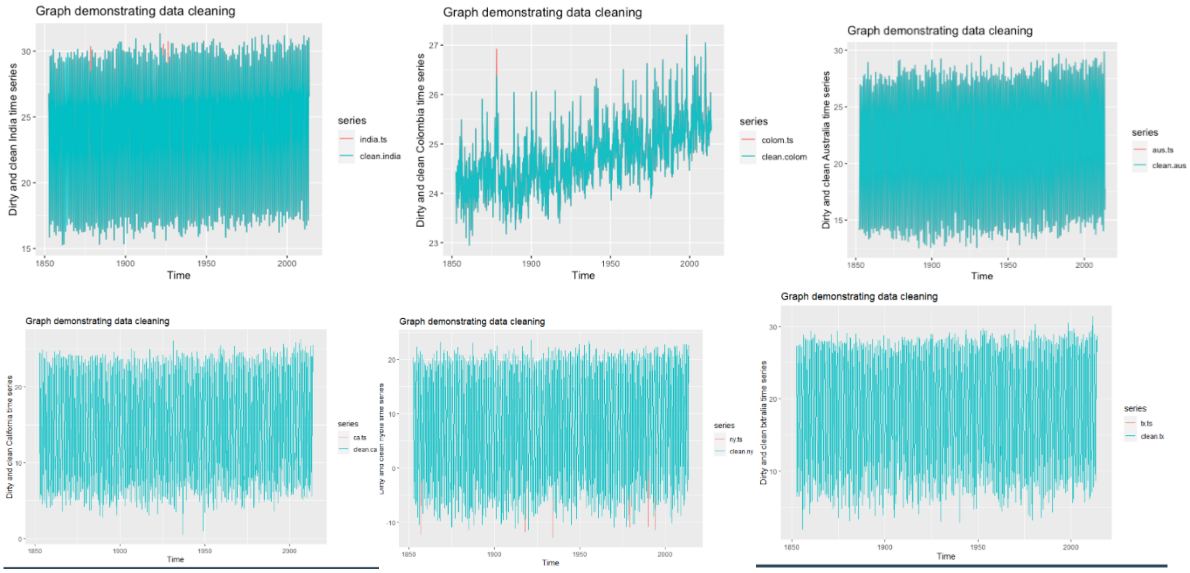

Figure 4. Comparison of before and after data cleaning

3.3 STL Decomposition

Given that we do not want to miss any values in the beginning and at the end caused by moving average feature in the traditional decomposition method and we would like to see the most recent value forecasted as well, so we decided to use the STL decomposition method.

We did not find that seasonal waves get bigger with the most recent time in all three countries and states in the original time series plots, so we set s.window=’periodic’ in the formula in R to decompose.

As shown below, all three countries have an upward trend but the trend in Colombia is more obvious. Plus, the forecast interval of Colombia in the trend panel is shorter than that in other countries but the forecast interval of Columbia in the seasonal panel is longer than others.

For the three states, plots still do not show a strong trend, and the trend or seasonality panel shows the same forecast interval among those three series.

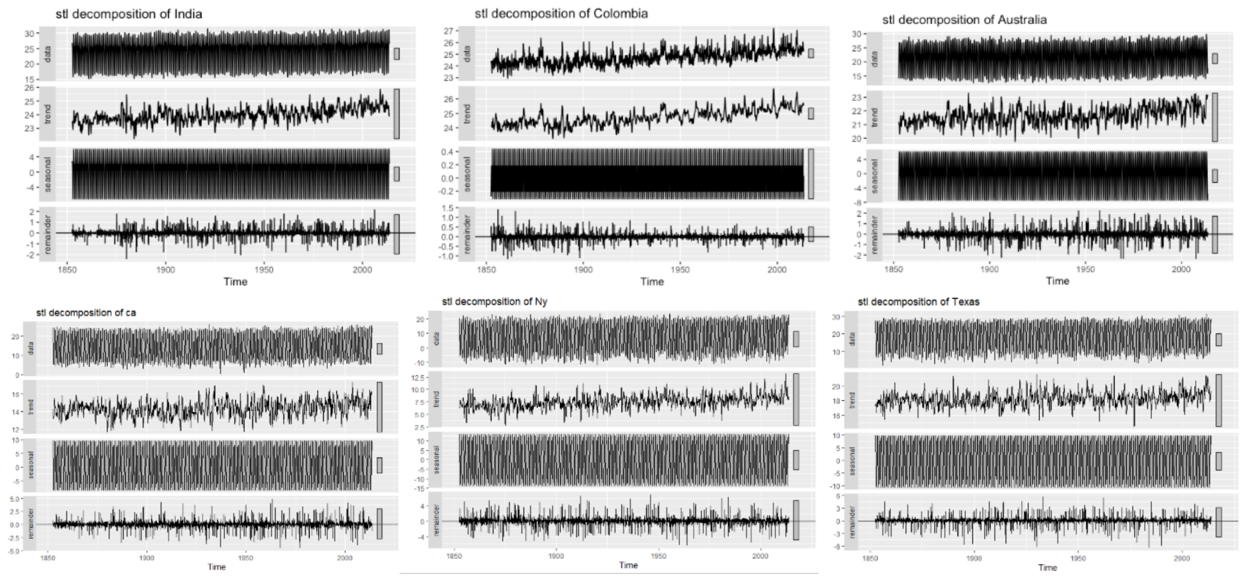

Figure 5. STL decomposition plots for countries and states

3.4 Strength of Trend and Seasonality

Here we'll use an objective method to calculate the strength of trend and seasonality that is only observing from plots we created before. As mentioned above, all three countries use the additive method, so the strength of the seasonality and trend should be accurate.

Table 1. Strength of trend and seasonality, Countries

Country	F_T	S_T
India	0.42	0.99
Colombia	0.77	0.41
Australia	0.36	0.98

Table 1. above verifies what we observed from Figure 5. before that Colombia has a strong trend but weak seasonality, both India and Australia have a strong seasonality but weak trend.

What shows in Table 2. below is also consistent with the findings we have from Figure 5., all those states show weak trends but strong seasonality.

Table 2. Strength of trend and seasonality, US States

State	F_T	S_T
California	0.20	0.96
New York	0.22	0.97
Texas	0.19	0.97

3.5 Unsupervised Tree Plot

The plot below shows that based on the similarity on values (Euclidean metric), India and Colombia are more similar, because both countries are in the North hemisphere, and Australia is in the South hemisphere. On the other hand, based on the similarity on the pattern (correlation metric), Colombia and Australia are closer.

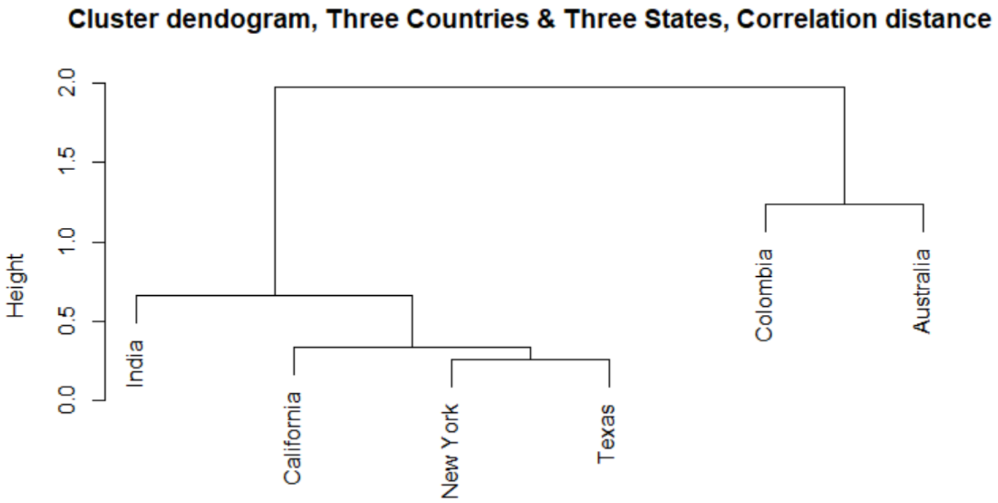

Figure 6. Correlation Cluster Dendrogram for countries and states

From the above tree in Figure 6, we can see that Colombia and Australia have branched together and had a similar pattern in their temperature over time. On the other branch, we can see that India has a similar pattern to the United States. Nevertheless, we can see that the states of New York and Texas are closely related, and California has a weaker correlation to India than it does to New York and Texas. This is expected as the three states are within the same country.

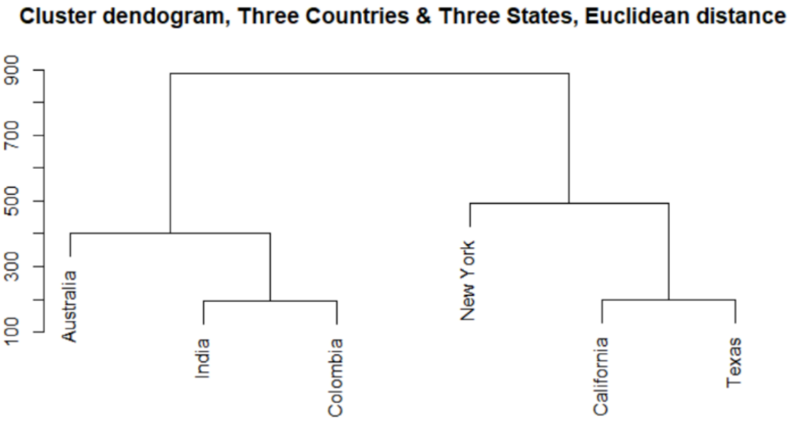

Figure 7. Euclidean Cluster Dendrogram for countries and states

In the above plot, the Euclidian distance method for clustering our data shows the United States separated from the other countries. California and Texas are closely related due to their high temperatures, and New York is colder but faces similar trends and seasonality to both states.

3.6 Entropy

We calculate the entropy values for all cleaned time series and find that Colombia has the largest entropy, which means that it is less predictable, and its time series is very chaotic. We will explore models in the modeling section to determine which model fits the best for each country and each state. Comparing the entropy values for the three states, we find that the cleaned California time series has the highest entropy.

Table 3. Entropy comparison in countries.

India	0.47
Colombia	1.68
Australia	0.58

Table 4. Entropy comparison in states

California	0.74
New York	0.65
Texas	0.69

4. Theory and Methodology

4.1 ETS

In ETS model, E is an error term, T is a trend component and S is a seasonal component. Parameters in this model are estimated by the method of maximum likelihood. In the output of R, when alpha is very big, xt and x(t+1) is very correlated (lag-1 plot will show a tight crowd on diagonal) and use equation below (xt is the next point at the mountain)

$$x_t = a_{t-1} + e_t$$

When alpha is very small, xt and x(t+1) is not correlated (lag-1 plot will show a scatter plot), and we use equation below (at is the next mountain level).

$$a_t = a_{t-1} + \alpha e_t$$

4.2 ARIMA (p, d, q)

The equation of AutoRegressiveMovingAverage can be written as below:

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p)(1 - B)^d X_t = c + (1 + \theta_1 B + \theta_2 B^2 + ...\theta_q B^q)\epsilon_t$$

‘d’ is the number of nonseasonal differences needed for stationarity (how many steps we can take to be stationary level), and ‘p’ is the number of autoregressive terms (how many past values in stationary level are used to forecast), ‘q’ is the number of lagged forecast errors in the prediction equation. If d = 0, R returns an estimate of μ; if d = 1, μ is returned as "drift" in the R output.

4.3 Neural Network

A Neural Network is one of the non-linearity methods, we do nonlinearity tests to see if the dataset is better suited to a non-linear model to fit by counting the number of p-values close to 0 out of 6.

4.4 Bagging

Bagging uses the bootstrap-aggregating method, and bootstraps similar series, and uses each of them to fit the ETS model and simulate one snapshot and average the resulting forecasts. We do not worry about bad metrics like higher MAPE and MASE shown in the result table and do not worry about the wider interval too, because bagging will consider three levels of uncertainties, even if one model makes mistake, the other one will cover it.

5. Modeling

5.1 Data Preparation (Data Cleaning: For India, Colombia, Australia, California, New York,)

- ❑ India: There are lots of missing values in the early years. We truncate the data so temperature data for India starts from Jan 1865.

- ❑ Australia: Temperature data for Australia starts from July 1852. We replaced 5 missing values using Tsclean().

- ❑ For the three states (NY, CA, TX) there were no missing values, and the same dataset was used as in the previous section.

  - India: There are lots of missing values in the early years. We truncate the data so temperature data for India starts from Jan 1865.
  - Colombia: Removed 200 continuous missing values before 1846 and one missing value in the end (September 2013).
  - Australia: Temperature data for Australia starts from July 1852. We replaced 5 missing values using Tsclean.
  - California, New York, and Texas same dataset

5.2 Train-test partitioning

For all three countries and three states, we choose the same train-test partitioning. The last three years' temperature data of each country and state are kept for testing.

5. 3 Model Building

In this section, we begin with a general comparative study among the best models from the ETS, SARIMA, and neural network categories. In addition, we conduct residual analyses checks on ETS and SARIMA models for each country and state. All test results indicate that the residuals are correlated, so some time dependence remains in the mistakes even after modeling. The non-linearity tests for each country and state imply that they all have a non-linear structure. Therefore, besides neural network model, we implement bagging as well. The accuracy metrics of the four competing models for the three countries and three states are listed in table 3 and table 5.

Table 5. Accuracy metrics for competing models, Countries

Country	Entropy	Fitted Model	Subset	AIC	MAPE	MASE
India	0.47	ETS (A, N, A)	Training	11215.78	1.952185	0.710652
			Testing		2.874747	1.019825
		ARIMA (2,1,3)	Training	6665.86	5.782149	2.117590
			Testing		24.132186	7.734308
		NNAR (28,1,14) [12]	Training		1.486339	0.5366606
			Testing		2.787691	0.9648472
		Bagging	**Training**		**1.858831**	**0.6740548**
		(15 boot, versions)	**Testing**		**2.914091**	**1.0488259**
Colombia	1.68	ETS(A,Ad,A)	Training	10619.38	1.015037	0.5885987
			Testing		1.564320	0.9260939
		ARIMA (0,1,3) (2,0,0)	Training	1505.68	1.085657	0.6304841
		[12]	Testing		2.464477	1.4578419
		NNAR (28,1,14) [12]	Training		0.5980821	0.3471332
			Testing		1.0864742	0.6472413
		Bagging	**Training**		**0.9422064**	**0.5466341**
		(15 boot, versions)	**Testing**		**1.8407460**	**1.0891489**
Australia	0.58	ETS (A, N, A)	Training	12962.39	2.662395	0.6978149
			Testing		4.232626	1.1239555
		ARIMA (5,1,0)	Training	6450.35	3.520145	0.9169788
			Testing		4.388209	1.1136704
		NNAR (32,1,16) [12]	Training		1.96502	0.5130779
			Testing		4.10004	1.0757641
		Bagging	**Training**		**2.547358**	**0.6683078**
		(15 boot, versions)	**Testing**		**4.388165**	**1.1644924**

To forecast Colombia's future average land temperature, we would implement the bagged model. The non-linearity tests suggest a non-linear model structure, which explains why the neural network has much smaller MAPE and MASE values on both the training and testing sets than the ETS and the SARIMA. However, as table 3 indicates, even the point accuracy on the test set of the bagged model is slightly worse than the neural network, but the neural network model has overfitting issue, it needs to recall 28 past values to forecast. In addition, the bagged model has taken care of the below three sources of uncertainty: model uncertainty, parameter uncertainty, and the random noise e_t.

Table 6. Accuracy metrics for competing models, States

Country	Entropy	Fitted Model	Subset	AIC	MAPE	MASE
California	0.74	ETS(ANA)	Training	15465.97	9.96	0.71
			Testing		8.12	0.72
		ARIMA (5,1,0)	Training	8811.75	16.38	1.39
			Testing		28.80	2.72
		NNAR (29,1,15) [12]	Training	-	7.07	0.50
			Testing		9.64	0.77
		Bagging (15 boot, versions)	Training		9.83	0.71
			Testing		8.14	0.74
New York	0.65	ETS (A, N, A)	Training	24650.56	Inf	0.70
			Testing		125.9054	0.67
		ARIMA (2,0,1) (2,1,0) [12]	Training	11249.87	Inf	1.16
			Testing		693.923	1.94
		NNAR (33,1,17) [12]	Training		6.67	0.53
			Testing		1435.47	5.79
		Bagging (15 boot, versions)	Training		Inf	0.69
			Testing		120.86	0.67
Texas	0.69	ETS (A, N, A)	Training	19137.69	8.63	0.71
			Testing		7.60	0.91
		ARIMA (1,0,1) (2,1,0) [12]	Training	8428.33	9.61	0.79
			Testing		7.17	0.86
		NNAR (27,1,14) [12]	Training		6.42	0.52
			Testing		9.48	1.23
		Bagging (15 boot, versions)	Training		8.87	0.70
			Testing		7.61	0.88

A bagged model was also used to forecast the land temperature of the states. As shown in the accuracy metrics table above, a bagged model was also used to forecast the land temperature of the states. As shown in the accuracy metrics table above, a Neural Network model performed better than the ETS and ARIMA models on MAPE and MASE values for both the training and testing set. We confirmed that the states require a nonlinear model using the non-linearity test and chose the bagged model as the final model due to its superior testing sample forecasting accuracies. This is due to the higher testing MAPE and MASE values of the Neural Network model despite it sometimes being the superior model for the training set, which shows some overfitting.

5.4 Forecasting through the best model

In this section, we use the best model from bagging approach to plot the temperature forecasting for all three countries and three states, please see Figure 8. ~ Figure 13. below:

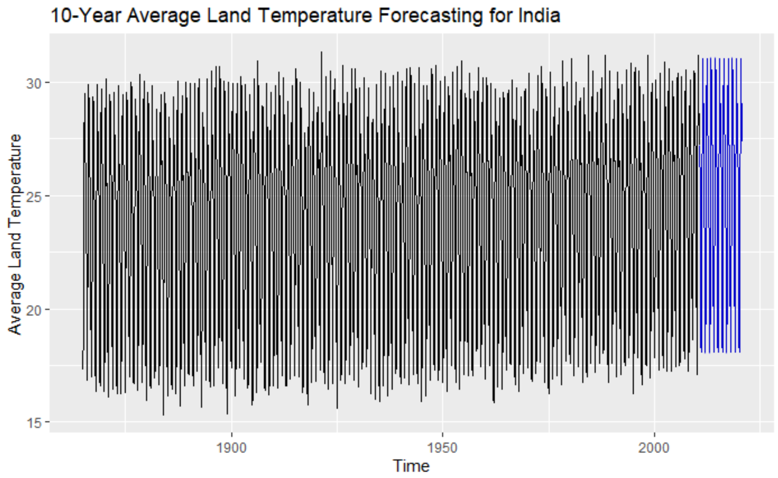

Figure 8. Forecasting for India through bagging

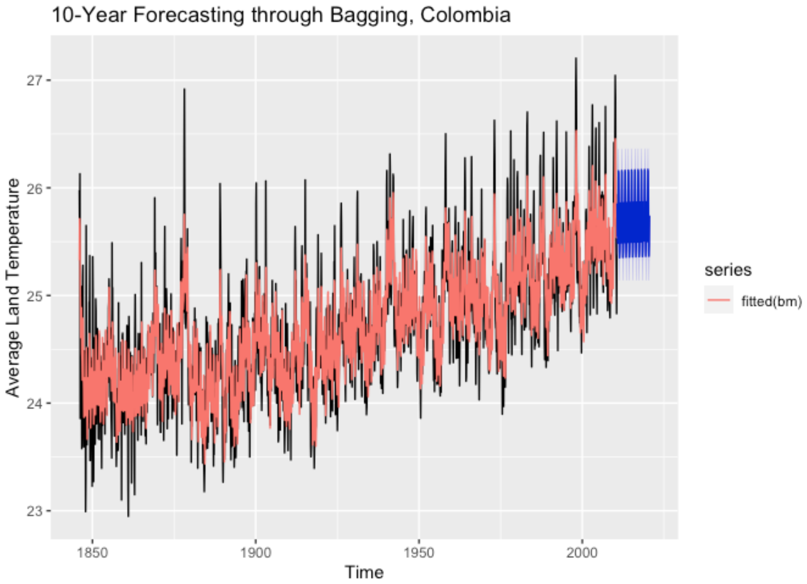

Figure 9. Forecasting for Colombia through bagging

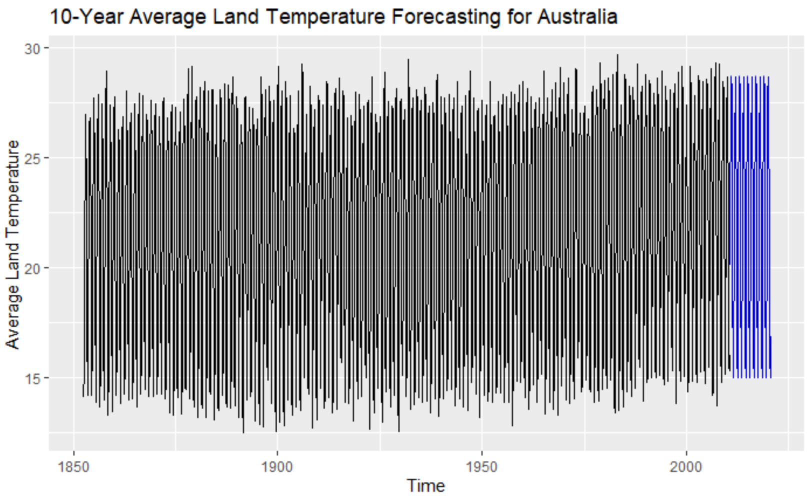

Figure 10. Forecasting for Australia through bagging

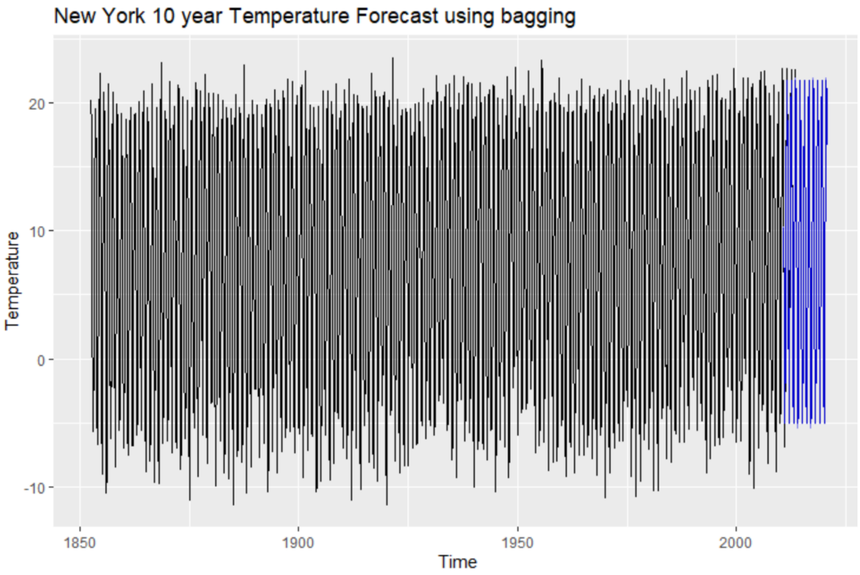

Figure 11. Forecasting for New York through bagging

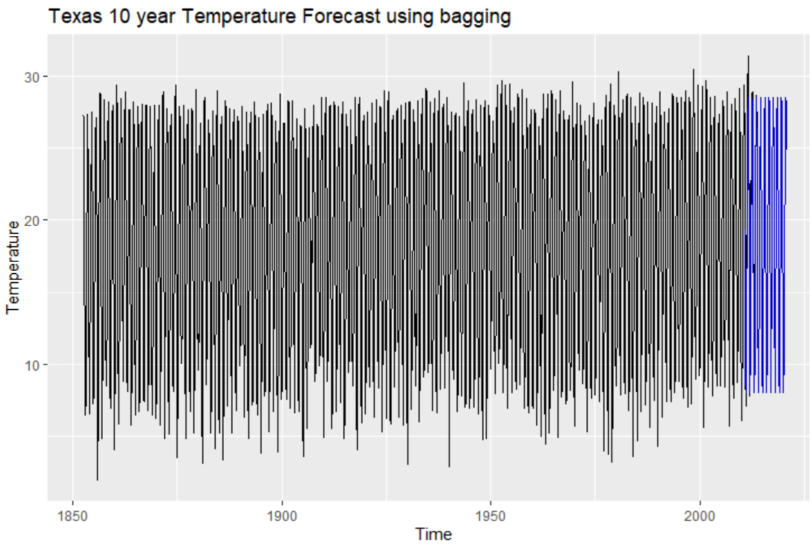

Figure 12. Forecasting for Texas through bagging

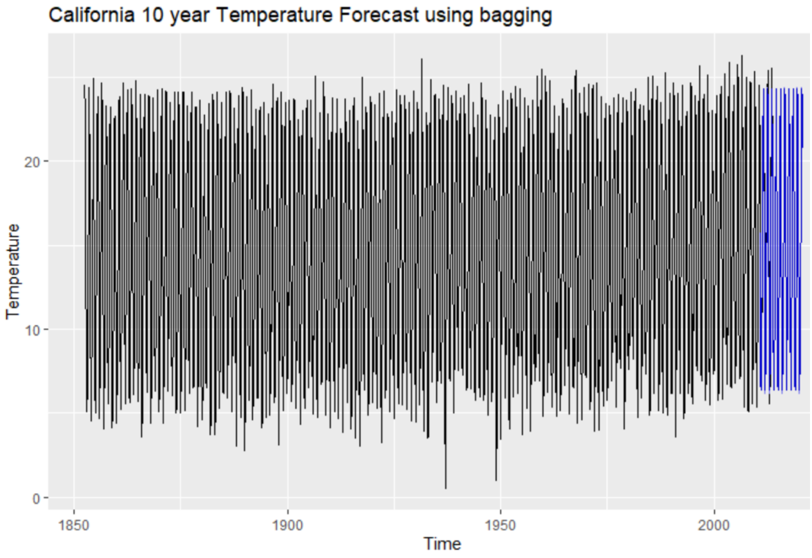

Figure 13. Forecasting for California through bagging

5.5 Temperature change comparison

To measure how much temperatures have changed for the three countries and three states, we calculate the difference between 2011 and 2020. We select two months January and July to represent the two seasons since each country may have different seasons depending on their locations. The results are showed in Table 7. below:

Table 7. Temperature change comparison for countries

India	Jan	July
2011	18.04191	28.02650
2020	18.05169	28.03593
Difference	+ 0.0978	+ 0.0943
Colombia	**Jan**	**July**
2011	25.78307	25.44722
2020	25.79731	25.46065
Difference	+ 0.1424	+ 0.1343
Australia	Jan	July
2011	28.65837	14.99751
2020	28.66759	15.00664
Difference	+ 0.0922	+ 0.0913

We repeat the same procedures for the three states, the results are showed in Table 8. below:

Table 8. Temperature change comparison for states

California	**Jan**	**July**
2011	6.389	24.306
2020	6.384	24.301
Difference	- 0.005	- 0.005
New York	**Jan**	**July**
2011	-5.017	21.77
2020	-5.001	21.79
Difference	+ 0.16	+ 0.02
Texas	**Jan**	**July**
2011	8.063	28.491
2020	8.057	28.486
Difference	- 0.006	- 0.005

Temperature Increase over the 9 years:

□ Colombia > India > Australia

□ New York > California > Texas

According to the results from Table 7. and Table 8. above, we could conclude that temperature change (and possible the climate change) is not uniform across the globe or even across one country. Some region oppositely experiences temperature decrease such as Texas. If we recall Texas's February cold wave that happened this year in 2021, we can say that it verifies our findings here.

6 Complementary Study

The greenhouse effect is closely related to global warming. Increasing emissions of greenhouse gases have caused natural warming of the earth since greenhouse gases in the atmosphere trap heat from the sun. Greenhouse gases include carbon dioxide, methane, nitrous oxide, water vapor, and fluorinated gases. Among them, carbon dioxide emissions are highly related to human activities and are often regarded as a symbol of the industrialization level of countries. In this section, we looked further into the annual CO2 emissions data to study whether CO2 emissions are related to the above countries' temperature changes.

Figure 14. shows the annual CO2 emission(tons) for India, the U.S., Australia, and Colombia. Overall, CO2 emission has been increasing in the past years. Even though the emissions decreased in a few years, that doesn't change the increasing trends. The left graph shows the comparison of the four countries. It clearly shows the huge difference in CO2 emissions across the countries. The U.S. has much more emissions compared to other countries. Colombia has a strong increasing trend in temperature; however, its CO2 emissions are the lowest.

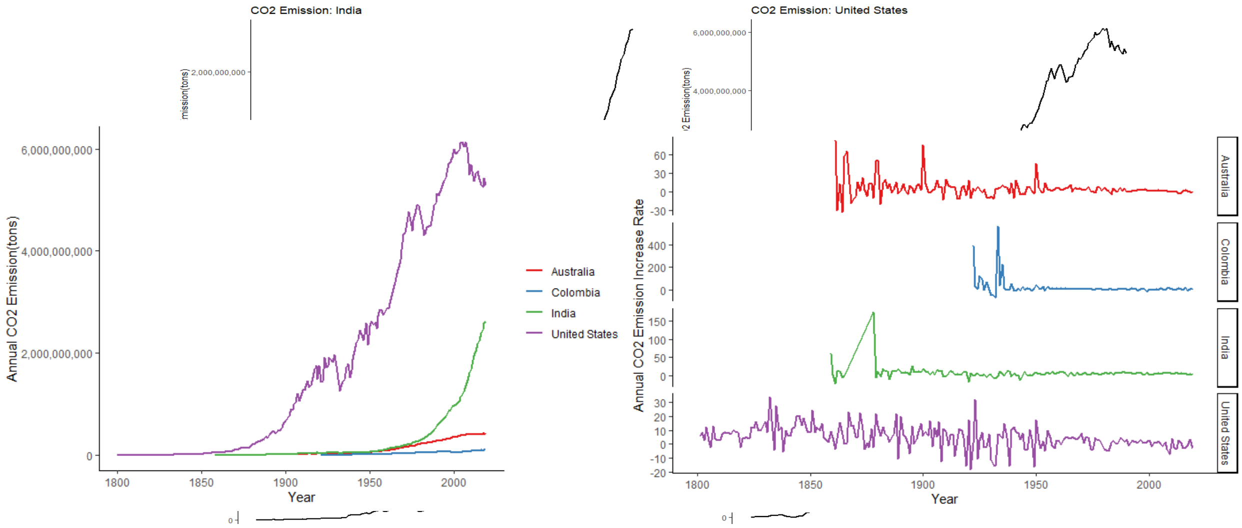

Figure 14 Comparison of Four Countries
Left: Annual CO2 Emission; Right: Annual CO2 Emission Increase Rate

The right graph shows the annual CO2 emission increase rate of the four countries. One significant pattern is that CO2 emission had increased rapidly before the middle 1900s and the rates have become flatter since that. Australia, India, and United States all have large increase rates in late 1800s to early 1900s, which is around the second industrial revolution. Colombia has extremely high increase rates in 1933 and 1935, which might be caused by its surge in industrialization that began in the 1930s.

7 Conclusion and future research

Our study suggests that temperature change is happening. Colombia shows a strong increasing trend in temperature. Even though other countries and states only show mild increasing trend, slight changes in average temperature matter. One research from NASA[1] explains why one degree increase in temperature has significant impact on the earth. Therefore, even tiny increasing trends are concerning.

Furthermore, climate change is a global issue. Every country is under the threat of extreme weather caused by global warming. A sad thing is that countries that do not produce many greenhouse gases will still be severely affected by climate change. Our time series analysis suggests that Colombia has a strong trend in temperature change. However, it produces much smaller CO2 compared to other countries. Therefore, it is crucial that the whole world works together to fight against climate change. Countries with high greenhouse gas emissions should take the lead to reduce emissions.

Our findings arouse our attention to actions to slower temperature changes. We would suggest that governments take active actions to encourage heavy industries to accelerate the transformation by adapting cleaner energy and renewable energy practices. In addition, markets can play a role in adjusting the industry structures, including agency acquisitions to foster markets for sustainable technologies and environmentally preferable materials, products, and services. Last but not least, collaboration among countries is extremely crucial.

Currently, there are some flaws in our study due to data and time limits, and we hope to improve it in the future. More data and variables should be included. We hope to investigate the temperature change from broader perspectives, such as changes in ocean temperature. Furthermore, more variables including monthly greenhouse gases emissions data, GDP, and industrialization level data can be added as regressors. We can also expand to more countries and cities around the world.

8 Reference

[1] https://www.nrdc.org/stories/greenhouse-effect-101#whatis

[2] https://climate.nasa.gov/

[3] https://climate.nasa.gov/news/2878/a-degree-of-concern-why-global-temperatures-matter/

[4] https://reliefweb.int/report/world/global-climate-risk-index-2021

[5] https://climate.nasa.gov/news/2878/a-degree-of-concern-why-global-temperatures-matter/

[1] https://climate.nasa.gov/news/2878/a-degree-of-concern-why-global-temperatures-matter/