Prediction of future global land temperature based on accuracy of different models and evaluating which model performs better

Pruthak Patel*, Rumana Myageri* and Zhimu Wang*

***Student at Harrisburg University of Science and Technology in the Analytics department**

ANLY -565 - Timeseries & Forecasting Final Project

Abstract

Global land temperature changes everyday with different rise and fall in various parts of the world, over the past 30 years, global land temperatures have increased to approx. 0.2°C per decade (James Hansen, 2006). This is no surprise that researchers predicted the rise of temperature in the early 1980s. Due to increased activities in the favor of global warming, several places have experienced strange climatic conditions in an decide. To name a few an article published in 2016 by researchers in Africa proved human activities brought drought in the region along with rising sea temperatures. (Pidcock, 2021). In this project we have studied global mean land surface temperature from 1970 to 2015 and used various models to identify which models provides the most accuracy with least mean absolute percentage error.

Introduction

Rising global land temperatures have brought several eye-opening scenarios for mankind. Such as increase in glacier melting, volcanic eruptions, shift of season by few weeks, scarcity of food and other resources that are directly or indirectly impacted by fluctuations in temperature. These temperature fluctuations become very important for growing vegetables, fruits, and other eating items, and therefore it is important to know which day/week/month of the year temperature will be highest and when it will be the lowest. In the present time many vegetables or food items are grown with the use of chemicals since farmers are unable to grow them in natural environment due to temperature fluctuations.

Using forecasting models that are available at our disposable that can be used to predict temperature for future, this can immensely help all those users whose work activities rely heavily on the temperature. To better evaluate each model and deeply understand its pros and cons, we will divide this project into four stages. In the first stage, we will clean and take aggregate of the data obtained from Kaggle. For the second stage, we will use models like autoregressive, regression, regression with ARIMA, Non-seasonal ARIMA, seasonal ARIMA, ARCH and VAR models. The second stage will help to answer research questions on which model performs better than other.

During third stage of the project, we will compare each model with others to identify the pros and cons of the model so when user must make an informed decision based on this project, this can provide substantial amount of information to do so. And, lastly, we will make conclusion on the use of the model, what worked best and why, and discuss any issues that we faced while doing the analysis of the available data. The use of any code during this project aligns with course requirements.

Research Question

1. Which model has the best prediction accuracy than others?
2. Which model has the least mean absolute percentage error?
3. What does the trend suggest for future? Will there by rise in temperature or fall and by how much?

Understanding Data

Source of the data is (Berkeley Earth Data). This data contains Global Land and Ocean-and-Land Temperatures (GlobalTemperatures.csv):

- Date: From 1750 for average land temperature and to 1850 for max and min land temperatures and global ocean and land temperatures
- LandAverageTemperature: Average temperature represented is in Celsius
- LandAverageTemperatureUncertainty: the 95% confidence interval around the average
- LandMaxTemperature: global average maximum land temperature in Celsius
- LandMaxTemperatureUncertainty: the 95% confidence interval around the maximum land temperature
- LandMinTemperature: global average minimum land temperature in Celsius
- LandMinTemperatureUncertainty: the 95% confidence interval around the minimum land temperature
- LandAndOceanAverageTemperature: global average land and ocean temperature in Celsius
- LandAndOceanAverageTemperatureUncertainty: the 95% confidence interval around the global average land and ocean temperature

Data Exploration

We will first explore the data to investigate any trends and seasonal patterns using autoplots and seasonal plots. For time series data, the obvious graph to start with is a time plot. That is, the observations are plotted against the time of observation, with consecutive observations joined by straight lines. Fig. 1. Shows the temperature autoplot for our dataset. This indicates there was a period after 1990 when temperatures peaked.

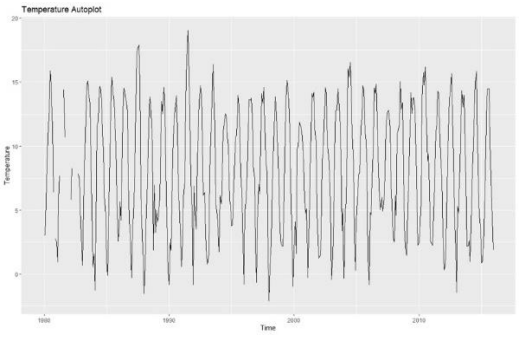

Figure 1: Temperature time plot from 1980 to 2015

A seasonal plot is like a time plot except that the data are plotted against the individual “seasons” in which the data were observed. Fig. 2 shows the temperature seasonal plot for our dataset, and clearly indicates rising temperatures in June and July every year.

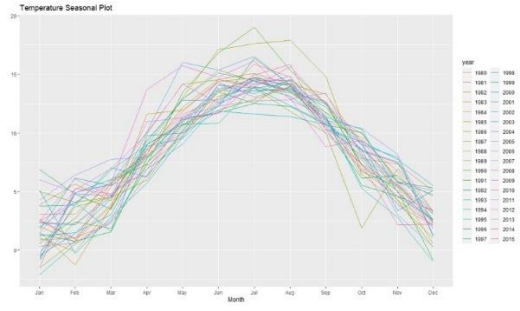

Figure 2: Seasonal Plot

A polar seasonal plot is a useful variation of the seasonal plot using polar coordinates. Setting polar=TRUE makes the time series axis circular rather than horizontal, as shown in Fig. 3.

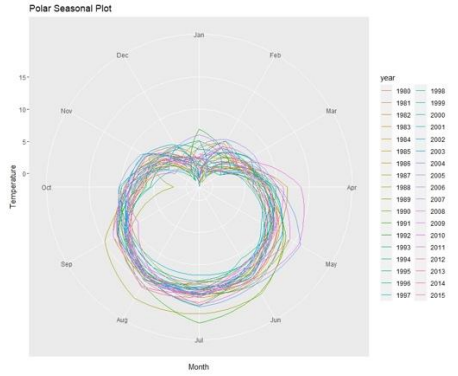

Figure 3: Polar seasonal plot

The times series plots of our data exhibit both seasonal and cyclic patters.
A subseriesplot is used to plot trend for every month separately, within the same plot. It is an alternative plot that emphasizes the seasonal patterns is where the data for each season are collected in separate mini time plots. The horizontal lines indicate the means for each month. This form of plot enables the underlying seasonal pattern to be seen clearly and shows the changes in seasonality over time. It is especially useful in identifying changes within seasons (Fig. 4)

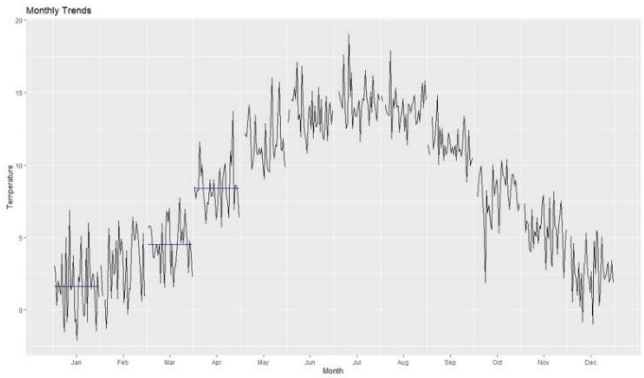

Figure 4: Monthly trends show Month 6 & 7 with the highest temperatures

A lag plot checks whether a data set or time series is random or not. Random data should not exhibit any identifiable structure in the lag plot. Non-random structure in the lag plot indicates that the underlying data are not random. For our dataset, the lagplots display distinct elliptical shapes. Single-cycle sinusoidal data gives rise to lag plots with circular or elliptical patterns. Values lying off the ellipse should be considered as potential outliers. Our sinusoidal data exhibits the typical elliptical pattern in the lag plots.

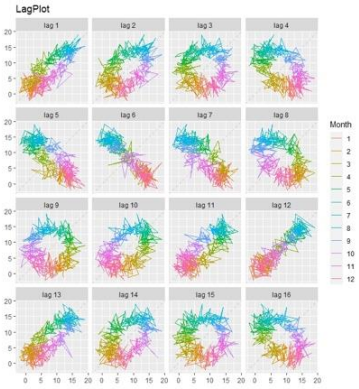

Figure 5: Lagplots exhibiting elliptical or circular patterns.

Fig. 6 shows the aggregate time series plot for our data without seasonal trends. It is evident that there is large fluctuation in temperatures from 19080 to 2015, with the highest temperatures being recorded in 2015.

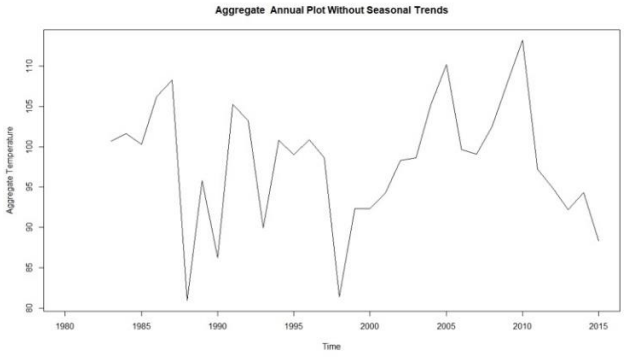

Figure 6: Aggregate annual plot without seasonal trends

We use the boxplot function to create monthly temperature boxplots for our data. The median is shown by the line inside the box of the boxplot. This may not always be in the middle – it depends on the shape of the distribution among other things. In Fig. 7 we see the mean temperatures rise

gradually and peak in months 6 and 7 during the summer, and then fall gradually again.

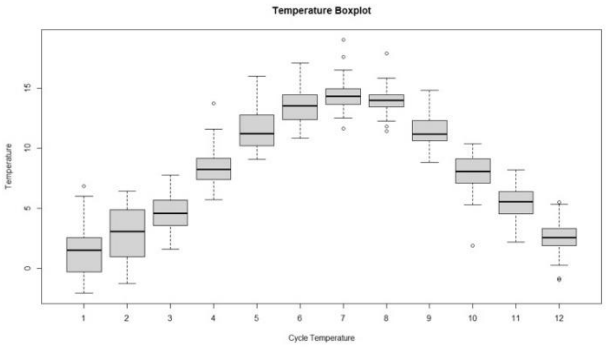

Figure 7: Monthly Temperature Boxplots

Autocorrelation

Autocorrelation plots (Box and Jenkins, pp. 28-32) are a commonly-used tool for checking randomness in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags. If random, such autocorrelations should be near zero for all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero.

In addition, autocorrelation plots are used in the model identification stage for Box-Jenkins autoregressive, moving average time series models. Note that uncorrelated does not necessarily mean random. Data that has significant autocorrelation is not random. However, data that does not show significant autocorrelation can still exhibit non-randomness in other ways. Autocorrelation is just one measure of randomness. In short, if the analyst does not check for randomness, then the validity of many of the statistical conclusions becomes suspect. The autocorrelation plot is an excellent way of checking for such randomness.Fig. 8 shows maximum autocorrelation occurring at lag 12. A distinct seasonal pattern is also observed which shows us our data is not random.

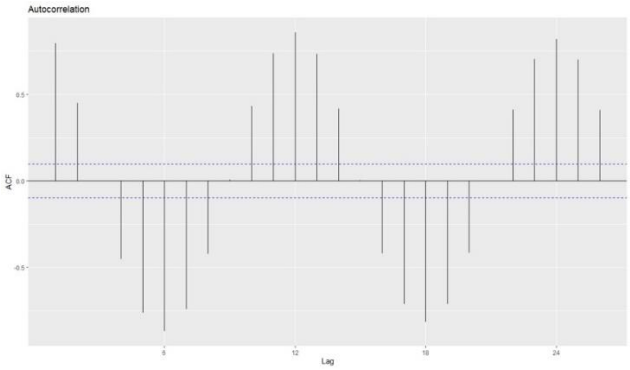

Figure 8: Autocorrelation Plot

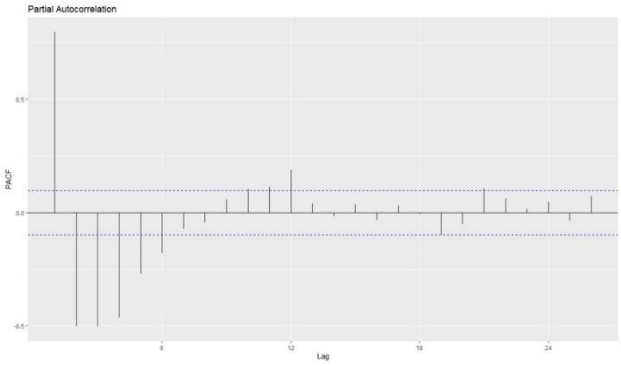

Figure 9: Partial Autocorrelation Plot

Holt Winters Smoothing

Holt (1957) and Winters (1960) extended Holt's method to capture seasonality. The Holt-Winters seasonal method comprises the forecast equation and three smoothing equations — one for the level ℓ_t, one for the trend b_t, and one for the seasonal component s_t, with corresponding smoothing parameters α, β^* and γ. We use mm to denote the frequency of the seasonality, i.e., the number of seasons in a year. For example, for quarterly data m=4m=4, and for monthly data m=12. In fig. 10, 2 models have been simulated, the first has no alpha and the second has alpha= 0.2. The sum of squared errors in the first model is 13.82 and 14.14 in the second model.

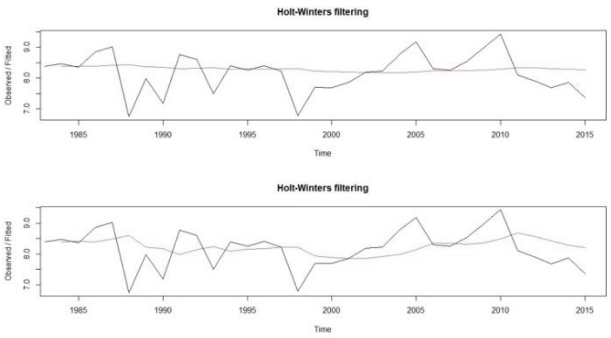

Figure 10: Holt Winters Smoothing

Forecasting using Seasonal Naïve Method

This method is like the naive method but **predicts the last observed value of the same season of the year**. Exponential smoothing uses the weighted average of previous observations where more weight is given to the most recent observations and the weight decreases as observations get older. Seasonal Naïve method is useful for highly seasonal data. In this case, we set each forecast to be equal to the last observed value from the same season of the year (e.g., the same month of the previous year). For example, with monthly data, the forecast for all future February values is equal to the last observed February value. With quarterly data, the forecast of all future Q2 values is equal to the last observed Q2 value (where Q2 means the second quarter). Similar rules apply for other months and quarters, and for other seasonal periods. The MAPE value for our forecast using this method is 65.65, which is too high.

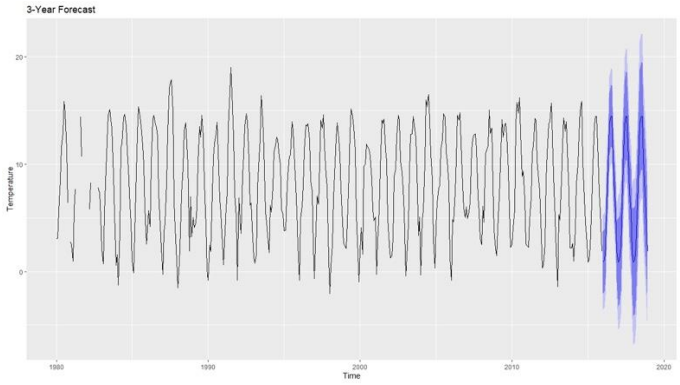

Figure 11:3-Year Forecast using Seasonal Naïve Method

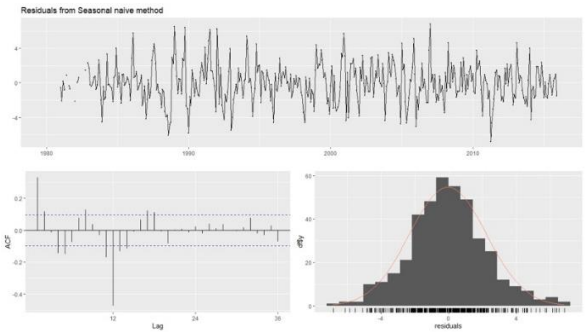

Figure 12: Residuals for Seasonal Naïve Method.

Autoregression

Autoregression is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step. It is a very simple idea that can result in accurate forecasts on a range of time series problems.

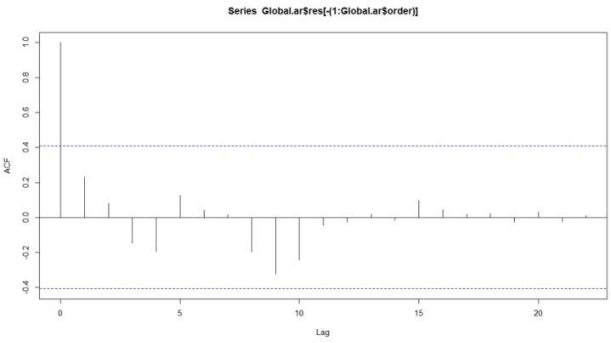

Figure 13: Autoregression (AR1) plot

Regression

There are two types of trends that can be observed in the regression model, stochastic and deterministic. Difference between the two models is while extrapolate deterministic trends are used to develop forecast, stochastic trends change the underlying trends slowly. It has been observed that there is significant difference between the time series regression and standard regression, in the situation when time series residuals correlation is positive the estimated standard error gives a value which is lesser than its true value. When we include the functions of time linear models are non-stationary. Figure 14 shows a regular seasonal

fluctuation which is shown using function decompose () in the main code file. Another thing to note in this graph is positive trend. The spread of seasonal fluctuations is present from 1970 to 2015.

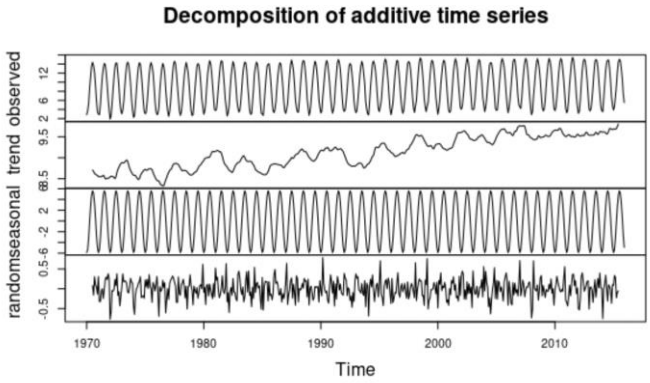

Figure 14: Seasonal fluctuation on global land temperature

Using maximum likelihood method, we were able to find the order of the model which is 12. During the analysis we found that if the aic () is set differently the order is different, but if we set it to false its 12. Figure 15 shows that the model does a pretty good job of showing seasonal variations.

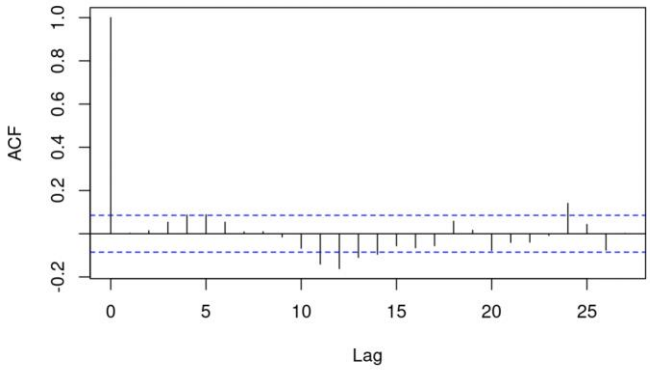

Figure 15: temp.ar$res[-(1:temp.ar$order)]

To determine whether the model is able to predict close to accurate values for temperature or not, we plot graph where red color shows observed values and blue to show forecasted values. Regression model is doing very good job in predicting the values which are very close to observed value, but important thing to note here is this graph is without the residuals of the forecasted.

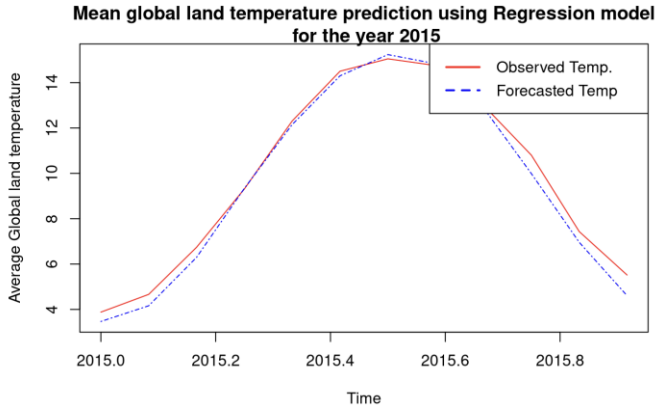

Figure 16: graph shows forecasted values vs observed values for the year 2015

Its important to calculate MAPE (Mean Absolute Percentage Error) this helps to answer our research question on whether this model provides better accuracy or not. MAPE for regression model is 5.34% which mean model is more than 92% accurate. With the residual series we will be able to determine more variation in prediction model. In the graph below it can be seen there is significant difference in the predicted value when we account for residuals and hence its MAPE is 37% which means model is not fit to be good fit.

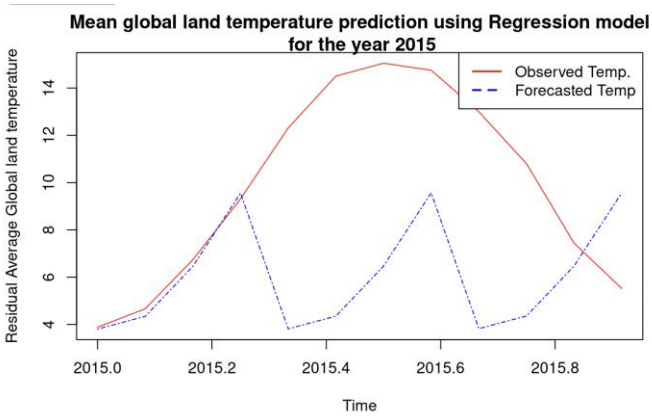

Figure 17: Residual series

Based on the graph 18 and 17 it can be concluded that regression model is able to predict better future values but only when not accounted

residuals, in the event when residuals are accounted it tends to deviate more from its accuracy.

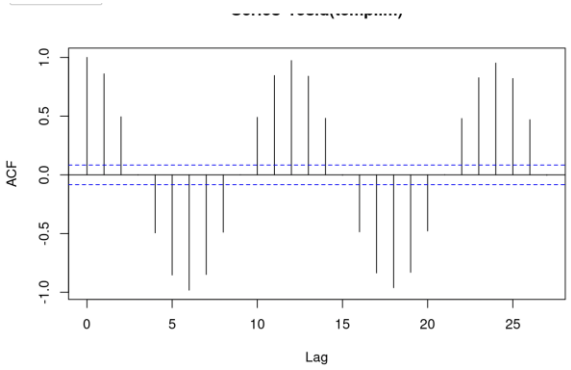

Figure 18: acf(resid(temp.lm))[1]

Regression + ARMA

The combination of regression and ARMA models can be used for various reasons such as to do simulation and fitting, predicting exchange rate series, electricity production series and many other. Here we use predict function that will be deployed to predict future values based on values obtained from the fitted regression model and most importantly able to calculate future errors that is attached to the forecasted regression model using ARMA process. (Buryi, 2021). The formulas that we will use for these models are:

Mean of ARMA(1,1) process

$$E(x_t) = 0$$

Variance of ARMA process

$$Var(x_t) = \sigma_w^2 + \sigma_w^2(\alpha + \beta)^2(1-\alpha^2)^{-1}$$

Autocorrelation of ARMA process

$$\rho_k = \frac{\alpha^{k-1}(\alpha + \beta)(1+\alpha\beta)}{1 + \alpha\beta + \beta^2}$$

Figure 19: Formulas of ARMA mixed models

The best order for ARMA model is found to be 1,0,1 without the seasonal component. The forecasted values of 12 months of 2015 are

```
Time Series:
Start = 541
End = 552
Frequency = 1
 [1] -0.0291821148 -0.0222168702 -0.0168714309 -0.0127691024 -0.0096207926 -0.0072046392 -0.0053503753
 [8] -0.0039273306 -0.0028352225 -0.0019970902 -0.0013538700 -0.0008602341
```

Table 1: Forecasted values of 12 months of 2015 using predict function

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
2015	3.785344	4.334114	6.456264	9.545105	12.348906	14.468474	15.313546	14.853686	13.058300	10.375312

	Nov	Dec
2015	7.133433	4.815623

Table 2: Forecasted values of 12 months of 2015 using time series function

Calculating the MAPE (Mean Absolete Percentage Error) for regression + ARMA model is 3.36% which means model is more than 95% accurate in predicting the better future values of the global mean land temperature. From Table below it can be observed that not all of the months shows same level of accuray in predicting the values for the future but only some which are in this case September, October and November. While for other months it cannot be said that model works well for them.

```
         2          3          4          5          6          7          8          9         10
13.538033  39.853359  39.097770  25.753897  15.840463   5.676570  -3.048972 -12.882411 -22.999481
        11         12
-37.463647 -39.292730
 [1]  18.378060  36.818648  32.335135  27.916312  16.391902   3.695102  -1.986243 -12.670958 -18.523371
[10] -37.370918 -29.791407
```

Table 3: Forecasted values and observed values of the year 2015

Looking at the two graphs below figure 20 and 21 it can be noted that when we look at the first graph the trends suggest the rate of temperature fluctuations has been significantly great in the early years before 2000s and after that change in temperature is not big, but when we look at the second graph, we can see that there are still significant fluctuations that occurs between 2000 to 2015. For us to see the seasonal trend in the temperature fluctuations, we can recommend that using ARMA model will be able to tell better story than other models since ARMA are widely used to see seasonal trends. With this we can also conclude about this model is that this model performs better than regression model due to its MAPE being lower than the other and able to provide users much better visual representation of the available dataset with seasonal fluctuations.

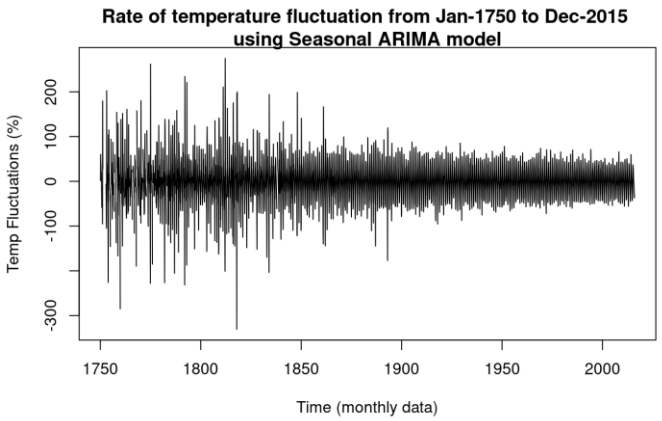

Figure 20: Rate of temperature fluctuations

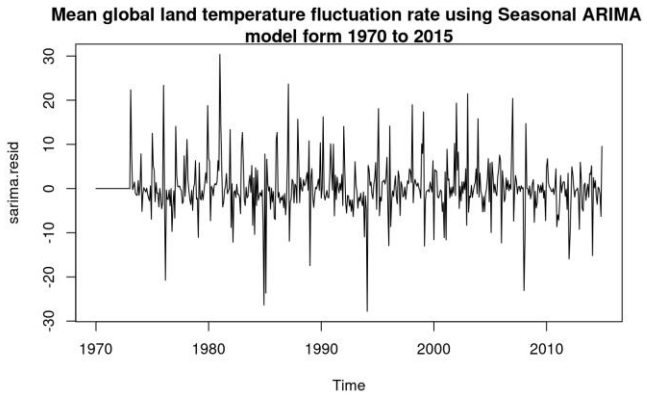

Figure 21: Seasonal ARMA trends

VAR

Vector Autoregressive model are considered to be stationary if the value of the determinant exceeds the value unity in absolute value (Buryi, 2021) The formula that we will use to for prediction of temperature fluctuations for the year 2015 from January to December are:

Two time series, $\{x_t\}$ and $\{y_t\}$, follow a vector autoregressive process of order 1 (denoted VAR(1)) if

$$x_t = \theta_{11}x_{t-1} + \theta_{12}y_{t-1} + w_{x,t}$$
$$y_t = \theta_{21}x_{t-1} + \theta_{22}y_{t-1} + w_{y,t}$$

where $\{w_{x,t}\}$ and $\{w_{y,t}\}$ are bivariate white noise and θ_{ij} are model parameters. Equation (11.2) can be rewritten in matrix notation as:

$$Zt = \Theta Z_{t-1} + w_t$$

Where

$$Zt = \begin{pmatrix} x_t \\ y_t \end{pmatrix}$$
$$\Theta = \begin{pmatrix} \theta_{11}\theta_{12} \\ \theta_{21}\theta_{22} \end{pmatrix}$$
$$w_t = \begin{pmatrix} w_{x,t} \\ w_{y,t} \end{pmatrix}$$

Figure 22: Formula to use VAR for prediction

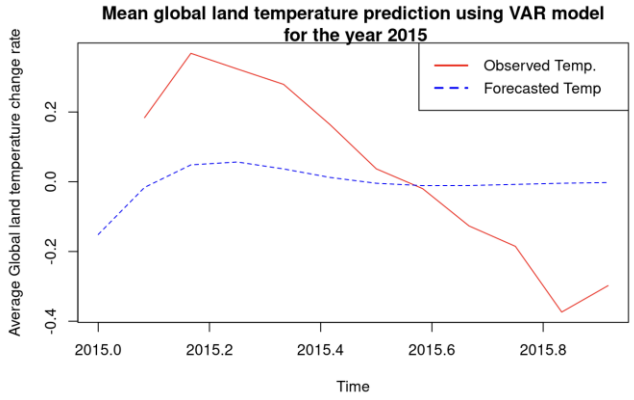

Figure 23: Observed value v/s predicted values using VAR model for 12-month prediction for the year 2015.

From above graph the forecasted values of the global mean land temperature deviates from its observed value and it may only be accurate for a month of June.

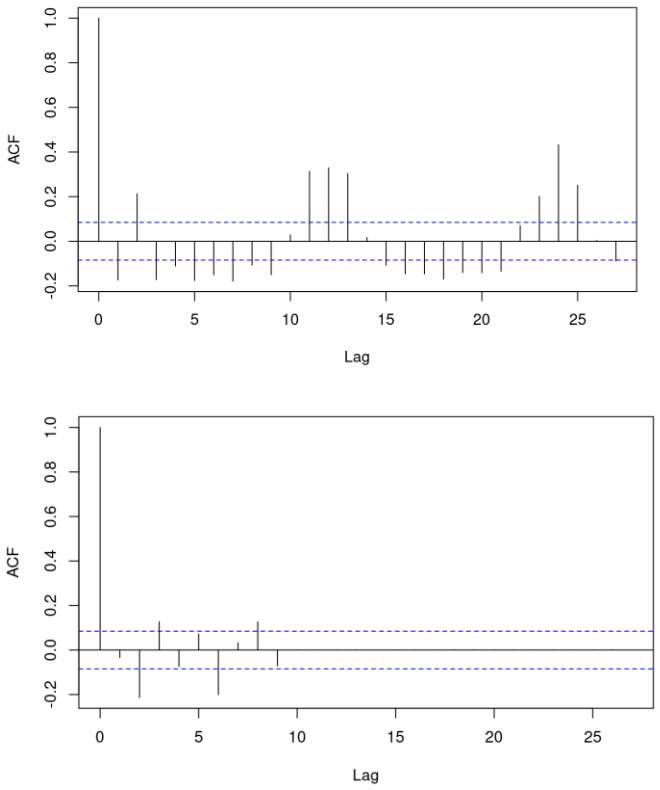

Figure 24: representation of no-autocorrelation

From the two graphs above there do not exist any autocorrelation between the predicted and

forecasted values of the model. Hence it can be said that VAR model does not fit very well for this dataset, further it can also be concluded that VAR model may fit for any other dataset that may not require to evaluate seasonal trends. A caviar to this model is that this model performs better to predict small range of future values compared to 12 month which involves 365 days of fluctuations. And hence VAR model is not a good fit for this dataset.

Non seasonal ARIMA

We also test how our data fits with non-seasonal ARIMA models. These models include autoregressive terms, moving average terms, and differencing operations. To determine the orders for the model (q: order of the moving average part; d: degree of first differencing involved; p: order of the autoregressive part), we use the ACF and PACF plots and manually select the best order which has the lowest AIC value. AIC value for each model we tried and summarized below.

Q,D,P	AIC
(4, 1, 1)	579.4807
(4, 1, 2)	427.9056
(4, 2, 2)	612.7501
(3, 0, 0)	696.9034
(3, 1, 0)	1088.121
(3, 1, 1)	652.887

Table 1: AIC of different Non seasonal ARIMA models

From the AIC value, we can see that the order (4, 1, 2) has the lowest AIC value which indicates best model fit to our data. We then use the model to forecast in 2015 and compare the actual 2015 temperature. The predictions are close to actual and the MAPE (Mean Absolete Percentage Error) is 3.78%

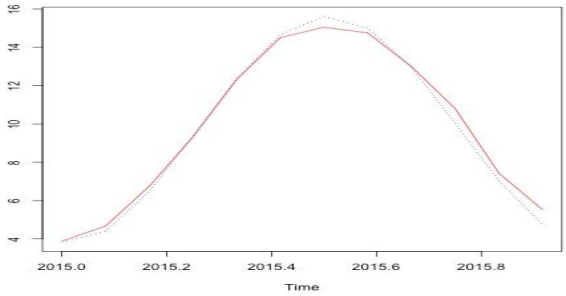

Figure 25: Prediction VS Actual for Non seasonal ARIMA model

Seasonal ARIMA

We then test how our data fits with seasonal ARIMA models. A seasonal ARIMA model uses differencing at a lag equal to the number of seasons (s) to remove additive seasonal effects. Seasonal ARIMA model consists following part:

- q: order of the moving average part;
- d: degree of first differencing involved;
- p: order of the autoregressive part;
- P: Seasonal autoregressive order;
- D: Seasonal difference order;
- Q: Seasonal moving average order;
- m: The number of time steps for a single seasonal period.

To find the best order, we use the similar function as shown in the class (get.best.sarima). We generate different pairs of orders by setting the max order, and then fit each order and compare the AIC value. Using this equation, we get the best order (2,1,2,2,1,2). Using the best model to predict on the 2015 data, MAPE (Mean Absolete Percentage Error) is 3.73%, slightly lower than the best Non seasonal ARIMA model.

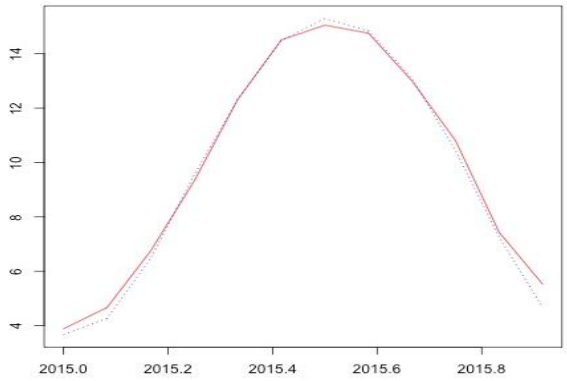

Figure 26: Prediction VS Actual for seasonal ARIMA model

Arch + GARCH

GARCH is used when the variance error is believed to be serially autocorrelated. GARCH models assume that the variance of the error term follows an autoregressive moving average process. We try the ***SARIMA + GARCH(1,1)*** for the temperature data. We first collect and inspect the residuals from the SARIMA model.

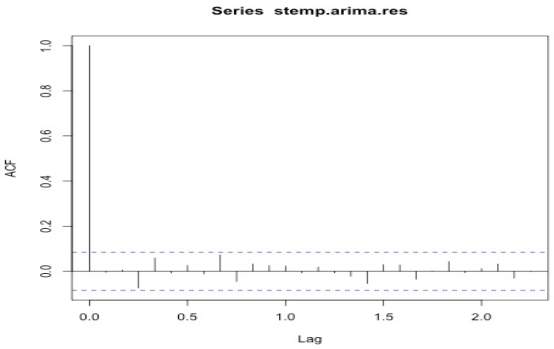

Figure 27: ACF for arima.res

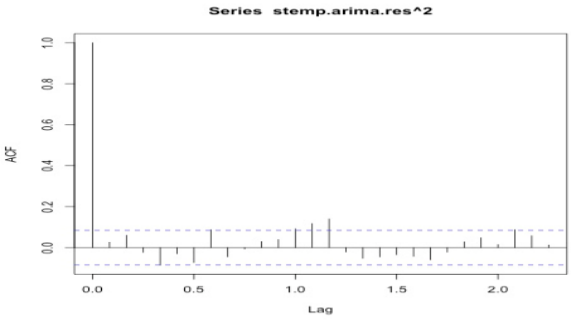

Figure 28: ACF for arima.res^2

From the ACF of residual and residual^2, we don't see clear autocorrelation. We still try to fit a garch model just to see how the result would look like. We can see that the meanError and standardDeviation goes up for later months in 2015. The residual variance grows bigger.

	meanForecast	meanError	standardDeviation
1	0	0.3563879	0.3563879
2	0	0.3669498	0.3669498
3	0	0.3778112	0.3778112
4	0	0.3889809	0.3889809
5	0	0.4004682	0.4004682
6	0	0.4122824	0.4122824
7	0	0.4244331	0.4244331
8	0	0.4369303	0.4369303
9	0	0.4497841	0.4497841
10	0	0.4630052	0.4630052
11	0	0.4766042	0.4766042
12	0	0.4905924	0.4905924

Table 2: GARCH Model Prediction

Conclusion

Seasonal ARIMA model and non-seasonal ARIMA model have very close mean absolute percentage error. Seasonal ARIMA model is slightly better than the non-seasonal ARIMA. The trend suggests Slow increase in global temperature.

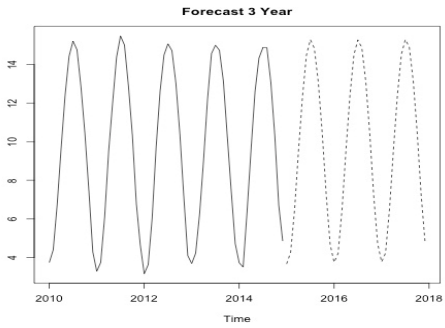

Figure 29: 3-Year Forecast using SARIMA

Year	Average
2010	9.703083
2011	9.516000
2012	9.507333
2013	9.606500
2014	9.570667
2015	9.693801
2016	9.713849
2017	9.738531

Figure 30 : 3- Year temperature forecasts

References

Bibliography

- Berkeley Earth Data. (n.d.). Climate Change: Earth Surface Temperature Data.
- Buryi, P. (2021, 09). Time Series and Forecasting model. PA, USA.
- James Hansen, M. S.-E. (2006). Global Temperature Change. *Proceedings of the National Academy of Sciences of the United States of America*, 103.
- P.D.Jones, T. a. (1986). Global temperature variations between 1861 and 1984. *Nature*, 430-434.
- Pidcock, R. (2021, 02 25). *Attributing extreme weather to climate change*. Retrieved from CarbonBrief: https://www.carbonbrief.org/mapped-how-climate-change-affects-extreme-weather-around-the-world