Analyzing Global Sea Level Change due to Global Warming and Forecasting Future Values

research project using time-series analysis

presented by:

Sanya Chauhan

Final Year Student, B.Sc. (Hons.) Statistics, LSR

ACKNOWLEDGMENT

I would like to convey my heartfelt gratitude to Dr. Renu Kaul for her tremendous support and guidance throughout this research project.

The completion of the project would not have been possible without her mentorship and insights.

INTRODUCTION

- ***WHAT DOES SEA LEVEL MEAN?***

it is the average height of the ocean surface.

- ***HAS THERE BEEN ANY CHANGE IN THIS SEA LEVEL?***

The warming of global climate has contributed significantly to a rise in mean sea level through two primary processes:

thermal expansion of water as it warms

the addition of freshwater to the ocean through the melting of terrestrial ice sheets and glaciers.

changes in land water storage that prevent water from entering the ocean (such as through agricultural use or the damming of water bodies) also impacts local sea levels.

HOW DOES IT IMPACT US

COASTAL LIFE

Serious threat of Flooding, Soil Erosion, and Tsunamis

Social Aspect: Displacement of millions of people living in coastal areas

LIFE FARTHER INLAND

Rising seas can contaminate soil and groundwater with salt.

Higher sea level causes heavy rains, strong winds and unleashes severe storms that can be a real threat to places that might be on its way

PURPOSE OF THE STUDY

The aim is to analyse change in global mean sea level in the past years and predict these values for the upcoming years

KEY CONCEPTS USED IN STUDY

- Stationarity of a Time Series Data Set
- Decomposition of Raw Data Set Using Ratio to Moving Average Method
- Mathematical curves
- ARIMA Modeling
- Ljung Box Test
- Forecasting

STATIONARITY

A stationary time series is one whose properties do not depend on the time at which the series is observed.

Conditions to prove that a Time Series is Weakly Stationary

- E[Yt] must be constant
- V[Yt] must be constant
- Cov[Yt, Y(t+k)] should only depend on lag k

Main Components which cause lack of Stationarity

- Trend
- Seasonality

ARIMA MODELS

Auto Regressive Integrated Moving Average Model is represented as ARIMA (p, d, q).

In practical life, data is generally not stationary and it becomes necessary to remove the non stationary sources of variation, i.e., trend and seasonality from the raw data before model fitting. ARIMA model is where non stationary component is removed by differencing the data one or more times.

General Equation:

$$W_t = \emptyset_1 W_{t-1} + \emptyset_2 W_{t-2} + \ldots + \emptyset_p W_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

DATA COLLECTION AND ASSUMPTIONS

1	Year	GMSL_noGIA	GMSL_GIA
2	1993	-38.59	-38.59
3	1993.027	-41.97	-41.97
4	1993.054	-41.93	-41.91
5	1993.081	-42.67	-42.65
6	1993.108	-37.86	-37.83
7	1993.135	-36.09	-36.05
8	1993.162	-36.11	-36.06
9	1993.189	-35.52	-35.47
10	1993.216	-35.47	-35.41
11	1993.243	-39.25	-39.19
12	1993.27	-37.52	-37.45
13	1993.297	-34.52	-34.45
14	1993.324	-36.63	-36.55
15	1993.351	-38.8	-38.71
16	1993.378	-40.23	-40.13
17	1993.405	-38.32	-38.21
18	1993.432	-36.03	-35.92
19	1993.459	-35.22	-35.1
20	1993.486	-35.38	-35.26
21	1993.514	-35.8	-35.68
22	1993.541	-28.46	-28.32
23	1993.568	-35.66	-35.52
24	1993.595	-34.13	-33.99
25	1993.622	-30.27	-30.11
1013	2020.324	45.89	52.59
1014	2020.351	47.19	53.88
1015	2020.378	48.94	55.65
1016	2020.405	49.79	56.51
1017	2020.432	50.71	57.44
1018	2020.459	49.87	56.61
1019	2020.486	53.62	60.38
1020	2020.514	55.82	62.56
1021	2020.541	56.61	63.34
1022	2020.568	55.56	62.35
1023	2020.595	57.64	64.39
1024	2020.622	57.24	63.99
1025	2020.649	54.87	61.59
1026	2020.676	57.25	64.1
1027	2020.703	56.2	62.98
1028	2020.73	52.07	58.88
1029	2020.757	52.56	59.34
1030	2020.784	50.1	56.88
1031	2020.811	51.24	58.03
1032	2020.838	52.6	59.39
1033	2020.865	52.7	59.5
1034	2020.892	52.74	59.54
1035	2020.919	53	59.8
1036	2020.946	49.05	55.87
1037	2020.973	47.1	53.93

GMSL: GLOBAL MEAN SEA LEVEL

GIA: GLOBAL ISOSTATIC ADJUSTMENT

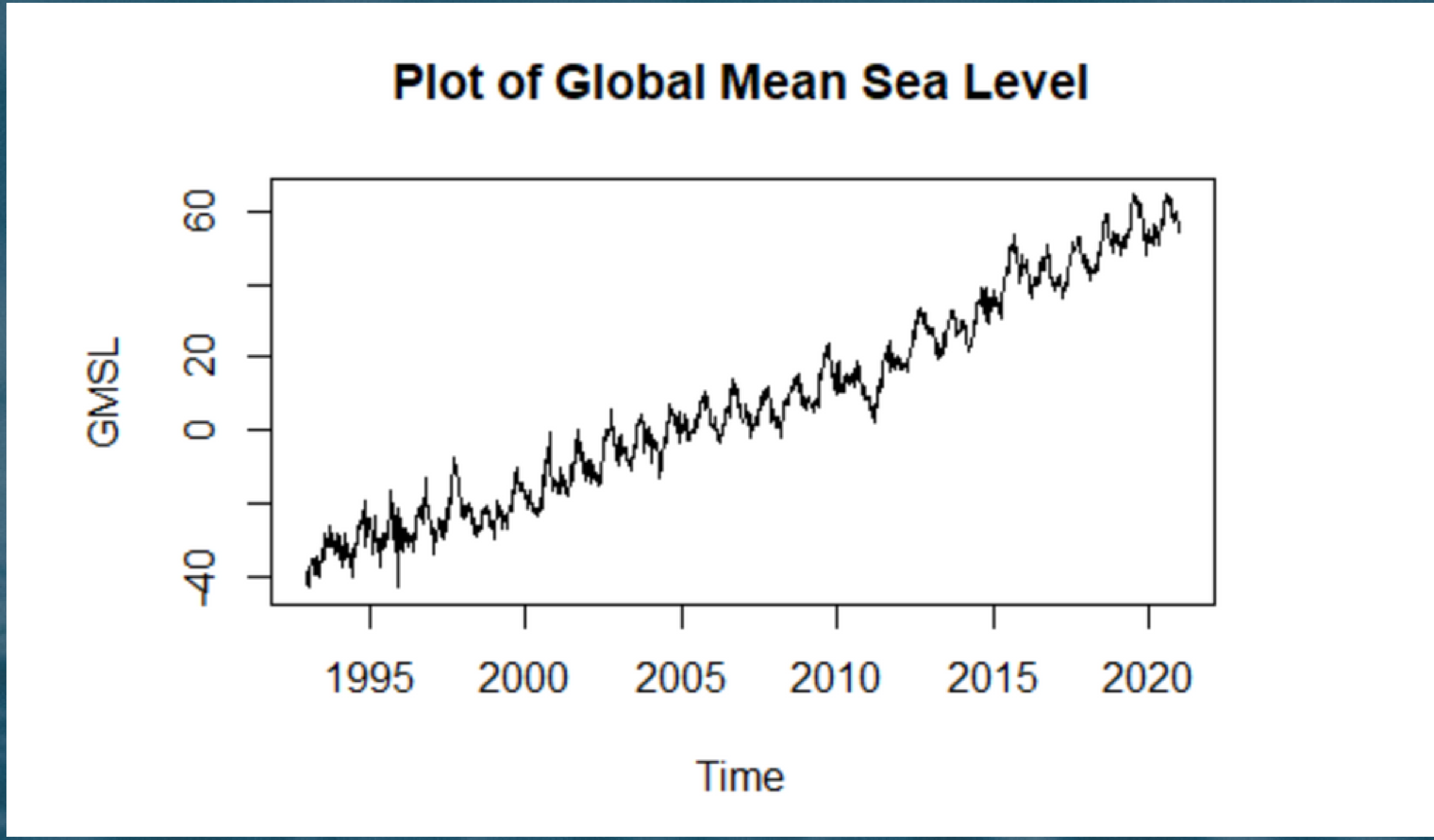
- This is secondary data collected from NASA's official website.
- The data was recorded by satellites and then interpreted and made public.
- The unit of the change in sea level is millimeter (mm).

ASSUMPTIONS

- For the data, it has been been assumed that the first observation is taken on 1st January 1993.
- While fitting the model, we assume that error terms follow normal distribution with mean 0 and a constant variance.

DATA ANALYSIS

SCATTER PLOT SHOWING RAW DATA



INTERPRETATION

This scatter plot is constructed to get an idea about the fit and stationarity of data. The scatter plot shows that data is not stationary since mean is not constant. This could be due to presence of an upward trend and seasonality in the data.

FITTING OF GROWTH CURVES

MODIFIED EXPONENTIAL CURVE

GMSL_GIA	First differences	Ratio of first differences
-38.59		
-41.97	-3.38	
-41.91	0.06	-0.0177515
-42.65	-0.74	-12.333333
-37.83	4.82	-6.5135135
-36.05	1.78	0.36929461
-36.06	-0.01	-0.005618
-35.47	0.59	-59
-35.41	0.06	0.10169492
-39.19	-3.78	-63
-37.45	1.74	-0.4603175
-34.45	3	1.72413793
-36.55	-2.1	-0.7
-38.71	-2.16	1.02857143
-40.13	-1.42	0.65740741
-38.21	1.92	-1.3521127
-35.92	2.29	1.19270833
-35.1	0.82	0.3580786
-35.26	-0.16	-0.195122
-35.68	-0.42	2.625

$$y_t = a + bc^t \text{ , } a > 0$$

As observed, since the ratio of the first differences of the consecutive values of yt is not constant. Thus, we cannot fit the modified exponential curve

GOMPERTZ CURVE

t	GMSL_GIA	Minimum of GMSL_GIA	GMSL_GIA + Constant	Fitting of Gompertz Curve by method of partial sums		Predicted GMSL_GIA
0.02703	-38.59	-43.14	5.55	S1=	7218.27	10554.3708
0.05405	-41.97		2.17	S2=	16754.53	10572.3485
0.08108	-41.91		2.23	S3=	28963.59	10590.3572
0.10811	-42.65		1.49			10608.3969
0.13514	-37.83		6.31	c=	1.0007206	10626.4676
0.16216	-36.05		8.09	B=	87.4143998	10644.5695
0.18919	-36.06		8.08	A=	-78.1518063	10662.7026
0.21622	-35.47		8.67			10680.867
0.24324	-35.41		8.73	b=	9.19584E+37	10699.0626
0.27027	-39.19		4.95	a=	1.14578E-34	10717.2896
0.2973	-37.45		6.69			10735.548
0.32432	-34.45		9.69			10753.8379
0.35135	-36.55		7.59			10772.1593
0.37838	-38.71		5.43			10790.5122
0.40541	-40.13		4.01			10808.8968
0.43243	-38.21		5.93			10827.3131
0.45946	-35.92		8.22			10845.7611
0.48649	-35.1		9.04			10864.2408
0.51351	-35.26		8.88			10882.7525

$$y_t = ab^{c^t}$$

As observed, the predicted values do not coincide (with some error) with the actual values. Thus, Gompertz curve is not a good fit to the data.

LOGISTIC CURVE

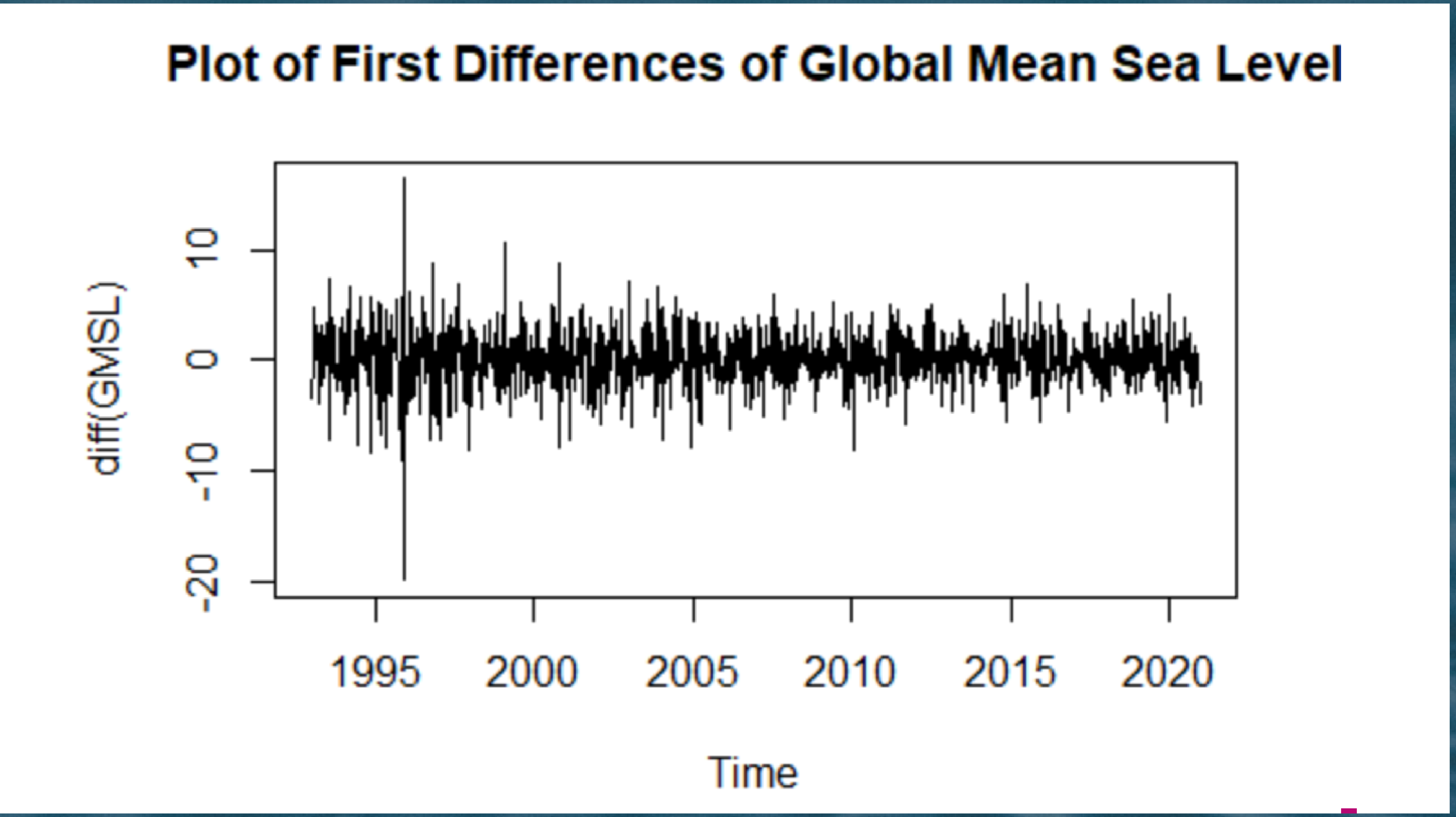
GMSL_GIA	Minimum of GMSL_GIA	GMSL_GIA + Constant	Reciprocal	First Differences	Ratio of First Differences
-38.59	-43.14	5.55	0.1801802		
-41.97		2.17	0.4608295	0.28064931	
-41.91		2.23	0.4484305	-0.012399	-0.04417969
-42.65		1.49	0.6711409	0.22271045	-17.9619687
-37.83		6.31	0.1584786	-0.5126623	-2.30192316
-36.05		8.09	0.1236094	-0.0348692	0.06801594
-36.06		8.08	0.1237624	0.00015298	-0.00438731
-35.47		8.67	0.1153403	-0.0084221	-55.0530565
-35.41		8.73	0.1145475	-0.0007927	0.09412313
-39.19		4.95	0.2020202	0.08747266	-110.345455
-37.45		6.69	0.1494768	-0.0525434	-0.60068332
-34.45		9.69	0.1031992	-0.0462777	0.88075157
-36.55		7.59	0.1317523	0.02855313	-0.61699605
-38.71		5.43	0.1841621	0.05240976	1.83551697
-40.13		4.01	0.2493766	0.0652145	1.24431976
-38.21		5.93	0.1686341	-0.0807425	-1.23810655
-35.92		8.22	0.1216545	-0.0469796	0.58184433
-35.1		9.04	0.1106195	-0.011035	0.23489006
-35.26		8.88	0.1126126	0.00199314	-0.18061964
-35.68		8.46	0.1182033	0.0055907	2.80496454

$$y_t = \frac{k}{1 + e^{a+bt}} \text{ , } b < 0$$

As observed, the ratio of the first differences of the consecutive values of reciprocals of yt is not constant, thus we cannot fit Logistic curve

DATA ANALYSIS

SCATTER PLOT SHOWING FIRST DIFFERENCES OF RAW DATA



INTERPRETATION

Plotting the first differences of the data to check for stationarity. It is observed that before 2000 the variance is high, however, after 2000 the variance is more or less constant. Thus, this plot shows that the data (could) be stationary. To confirm this belief , appropriate tests for stationarity of data are performed further.

STATISTICAL TESTS TO CHECK STATIONARITY OF DATA

Carrying out 3 tests for testing stationarity: PP.test, adf.test and kpss.test

```
#Testing Stationarity of the Data

#PP.test

PP.test(GMSL)
#p-value = 0.01
#Hence H0 is rejected

PP.test(diff(GMSL))
#p-value = 0.01
#Hence H0 is rejected

#adf.test

adf.test(GMSL)
#p-value = 0.01
#Hence

adf.test(diff(GMSL))
#p-value = 0.01

#kpss.test

kpss.test(GMSL)
#p-value = 0.01
#Hence not stationary

kpss.test(diff(GMSL))
#p-value = 0.01
#Hence not stationary
```

R CODE FOR TESTING STATIONARITY

SETTING HYPOTHESIS

For PP.test:

H0: The time series has a unit root

H1: There is no unit root

For adf.test:

H0: unit root is present in a time series sample

H1: Absence of unit root

For kpss.test:

H0: The univariate time series is trend stationary

H1: The time series is not stationary

RESULTS OF TESTING

It can be observed from PP.test and adf.test that the data is stationary. However, the scatter plot clearly shows lack of stationarity in the data. **Additionally, it is known that in presence of auto-correlation in the data, adf.test and PP.test give misleading results. Since there is presence of Autocorrelation in the given data, it is best to disregard the the inferences from both the above tests.**

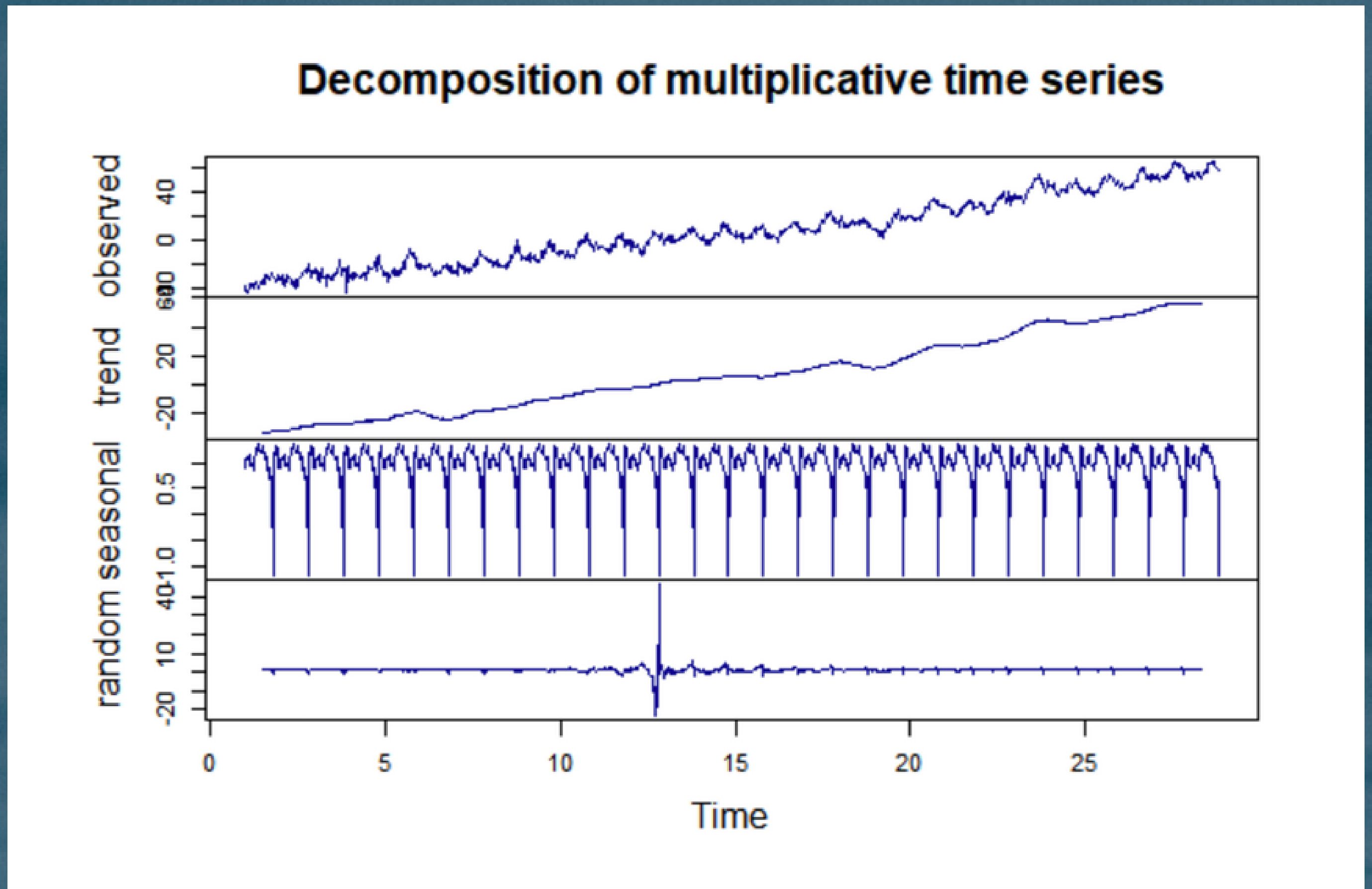
Hence, kpss.test gives the **most valid result** showing that the data is not stationary, and proves to be better than pp.test and adf.test

R CODE FOR TESTING SEASONALITY

```
tsgmsl <- ts(GMSL,frequency = 37,start = c(1993,1),end=c(2020,37))

library(seastests)

isSeasonal(tsgmsl)
#TRUE

isSeasonal(diff(tsgmsl))
#TRUE
```

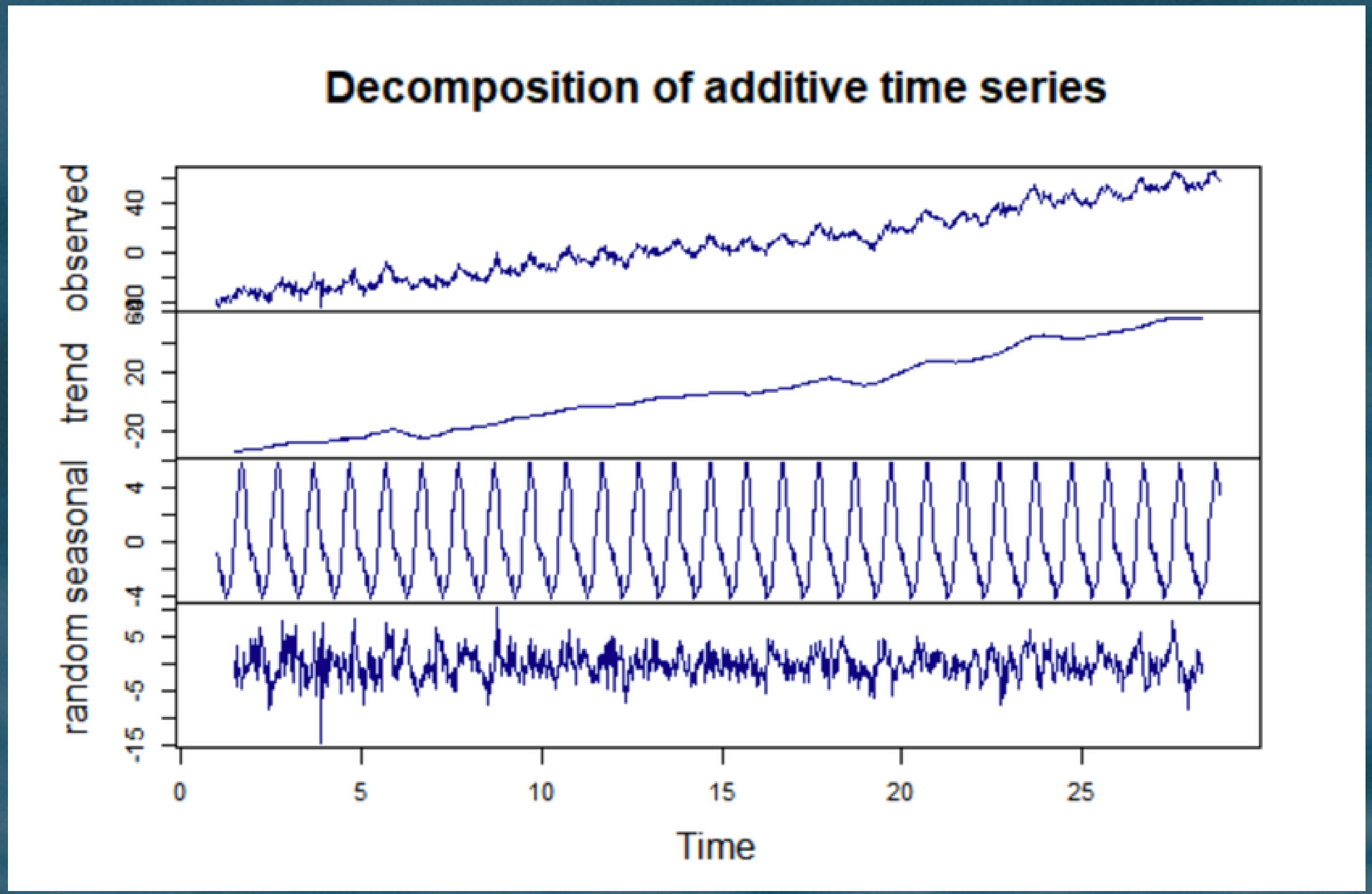
Checking for the second cause for non-stationarity, i.e., the presence of seasonality. It can be seen that the data is seasonal.

DECOMPOSITION OF DATA



Decomposing the time series into trend, seasonal and random component, for both, an additive and multiplicative model.

```
desgmsl <- decompose(tsgmsl, type = "additive")
plot(desgmsl, col="navy blue")

desgmslm <- decompose(tsgmsl, type = "multiplicative")
plot(desgmslm,col="dark blue")
```



It can be observed that in the multiplicative model the random component doesn't appear to be random upon observing the graph, hence, we reject the multiplicative model and go ahead with the additive model. Additionally, the additive model is useful when the seasonal variation is relatively constant over time and the multiplicative model is useful when the seasonal variation increases over time. Hence, the additive model is selected for the further analysis.

FITTING OF TREND COMPONENT

Extracting the trend component by subtracting moving average of extent 37 from original values. Further, fitting various curves using IBM SPSS

LINEAR

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.990	.981	**.981**	3.680

QUADRATIC

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.995	.990	**.990**	2.591

CUBIC

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.996	.991	**.991**	2.479

LOGARITHMIC

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.880	.775	**.775**	12.551

The independent variable is Time.

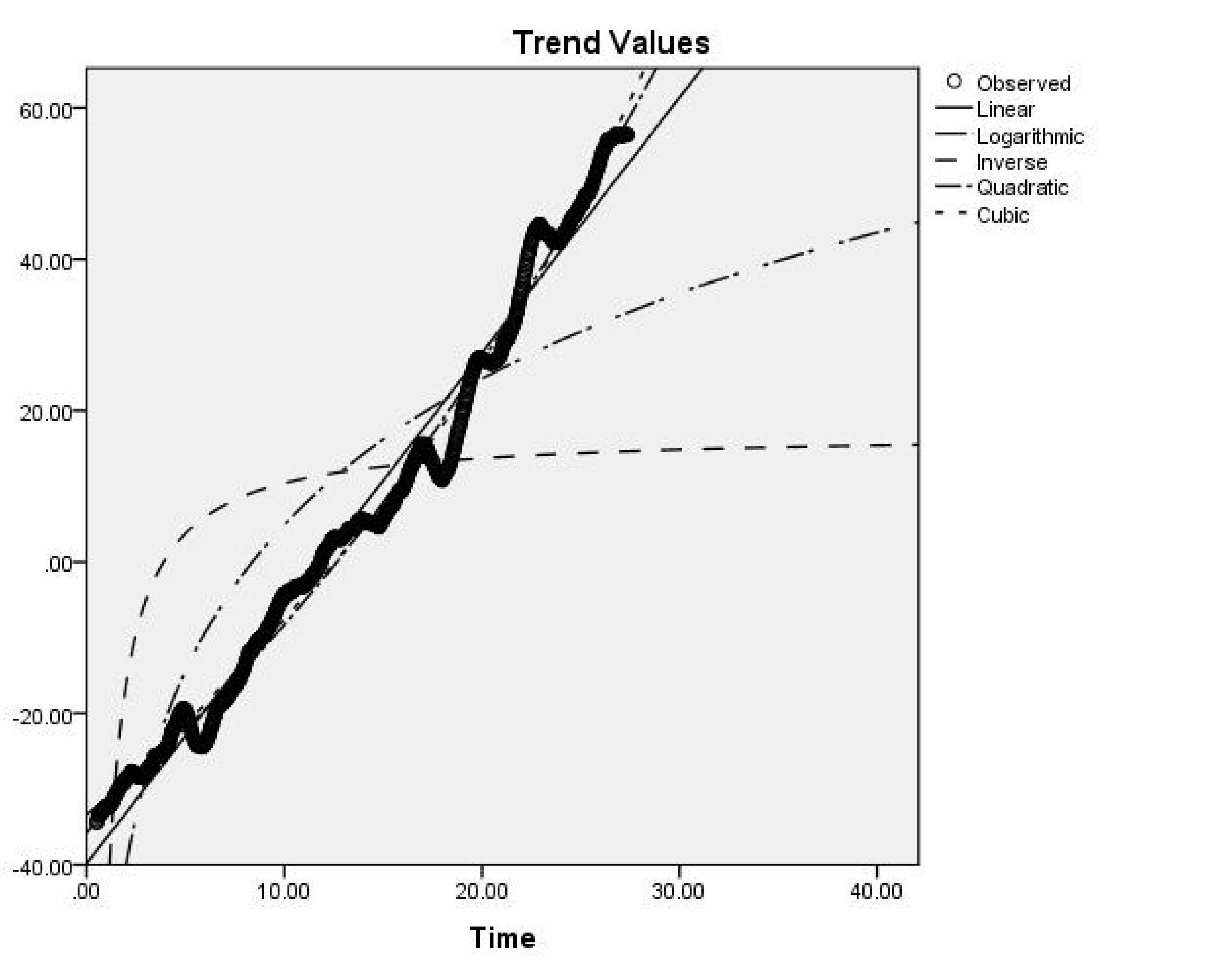
INVERSE

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.568	.322	**.322**	21.778

The independent variable is Time.

The values of Adjusted R^2 for Linear, Quadratic, and Cubic curves are very close. Hence, it calls for further analysis.



FITTING OF TREND COMPONENT

1. LINEAR

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	680796.838	1	680796.838	50277.443	.000
Residual	13432.474	992	**13.541**		
Total	694229.313	993			

2. QUADRATIC

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	687577.316	2	343788.658	51216.887	.000
Residual	6651.997	991	**6.712**		
Total	694229.313	993			

3. CUBIC

ANOVA

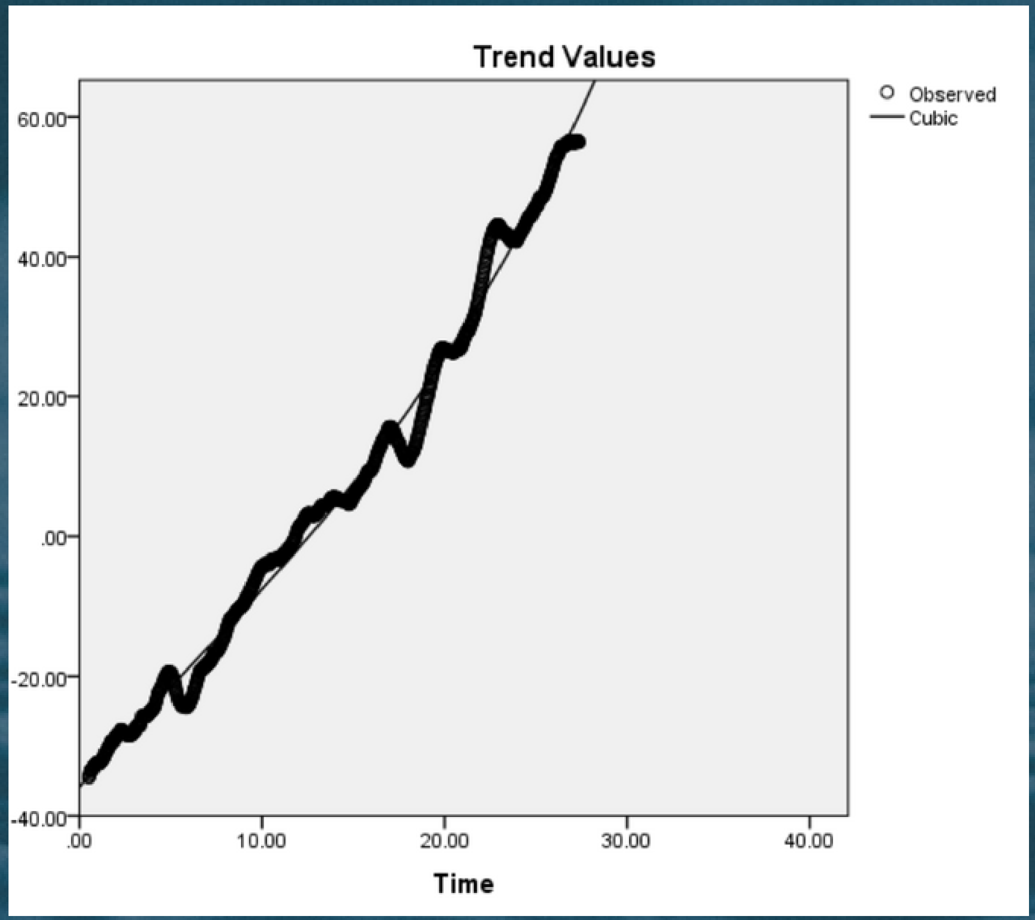
	Sum of Squares	df	Mean Square	F	Sig.
Regression	688142.889	3	229380.963	37310.442	.000
Residual	6086.424	990	**6.148**		
Total	694229.313	993			

On comparing the Mean Square Error, it can be noted that the cubic curve had the least value. Therefore, it is the best fit for the data.

FITTED CURVE: CUBIC

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Time	2.998	.110	.880	27.233	.000
Time ** 2	-.037	.009	-.316	-4.125	.000
Time ** 3	.002	.000	.453	9.591	**.000**
(Constant)	-35.848	.361		-99.263	.000



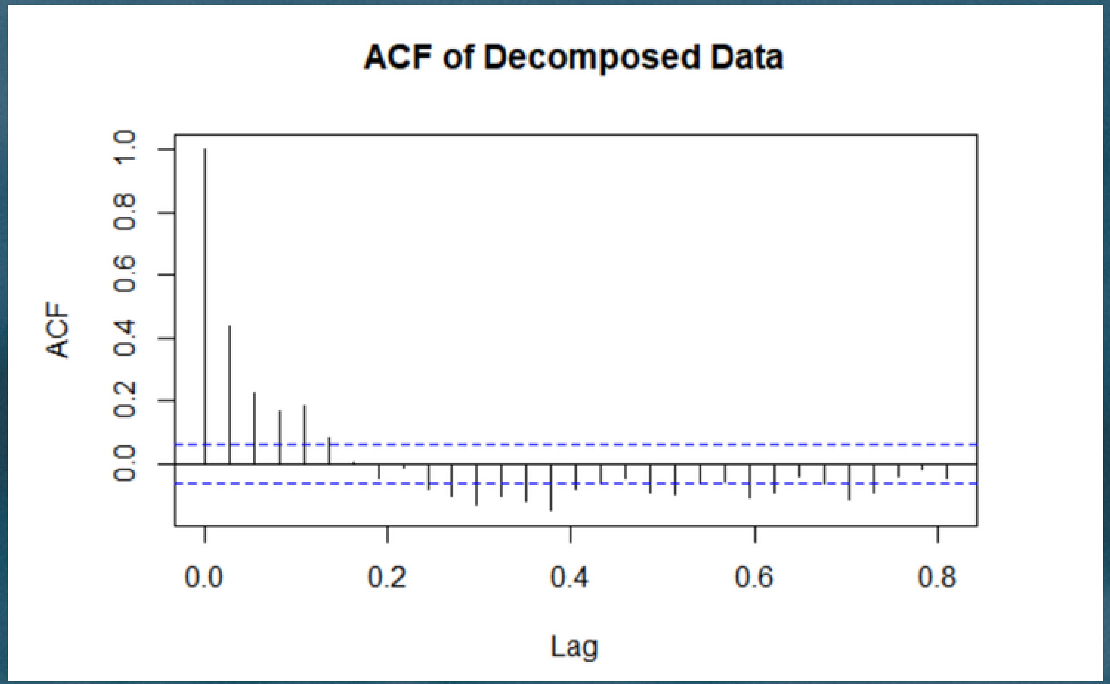
FITTING OF THE STATIONARY COMPONENT

Extracting the stationary component and checking its head & tail

```
rand.ser <- desgmsl$random  #Extracting the random component
head(rand.ser) #Checking the head of the random component
#[1] NA NA NA NA NA NA

tail(rand.ser) #Checking the tail of the random component
#[1] NA NA NA NA NA NA

#Head and tail contains NA since the initial and last k(=18) observations
#have been lost due to moving average of extent m = 2k + 1 =37
```

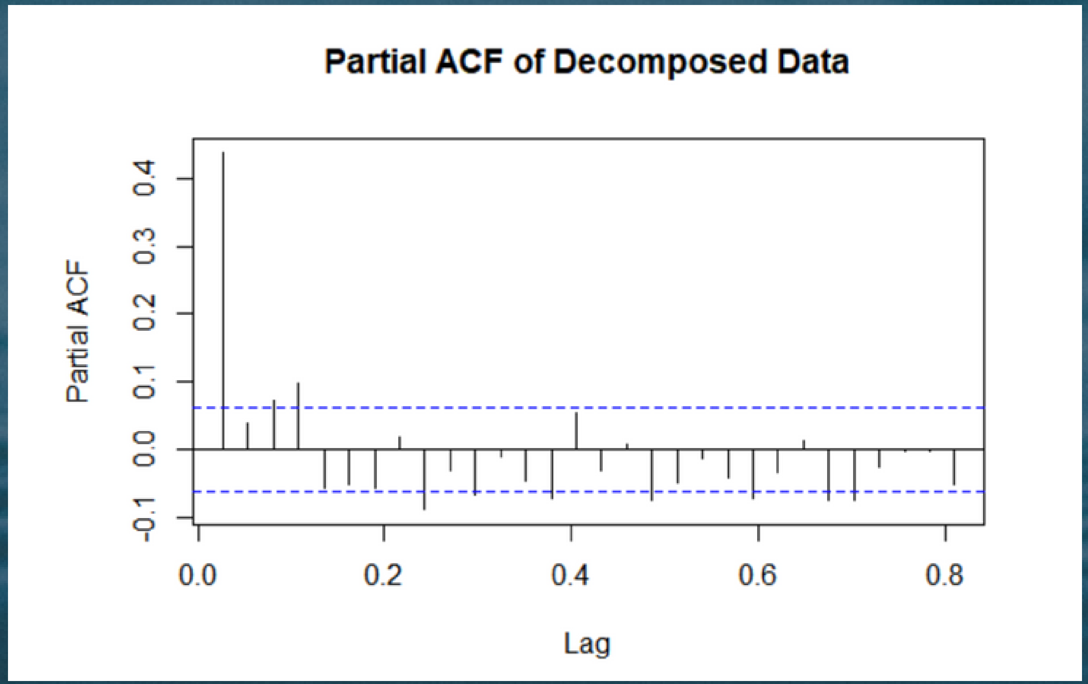
Using ran.series function to extract non NA values and concluding that the data is not seasonal.

```
ran.series <- rand.ser[19:1018] #Extracting non NA values

ran.ts <- ts(ran.series,frequency = 37,start=c(1993,19),end=c(2020,19))

isSeasonal(ran.ts)
#FALSE
```



Plotting their ACF and partial ACF and observe that the data is not seasonal and there is presence of geometric decay. Hence, it indicates that is an AR MODEL

Autocorrelation Function and Partial Autocorrelation Function Graphs of the Decomposed Data



MODEL FITTING

```
fit1 <- arima(ran.ts,order=c(1,0,0))
fit2 <- arima(ran.ts,order=c(1,0,1))
fit3 <- arima(ran.ts,order=c(1,0,2))
fit4 <- arima(ran.ts,order=c(2,0,0))
fit5 <- arima(ran.ts,order=c(2,0,1))
fit6 <- arima(ran.ts,order=c(2,0,2))
fit7 <- arima(ran.ts,order=c(0,0,1))
fit8 <- arima(ran.ts,order=c(0,0,2))
```

Fitting ARIMA Model to the decomposed data and extracting the residuals for all the fitted models

```
res1 <- fit1$residuals
res2 <- fit2$residuals
res3 <- fit3$residuals
res4 <- fit4$residuals
res5 <- fit5$residuals
res6 <- fit6$residuals
res7 <- fit7$residuals
res8 <- fit8$residuals
```

LJUNG BOX TEST TO TEST THE FIT OF THE MODEL

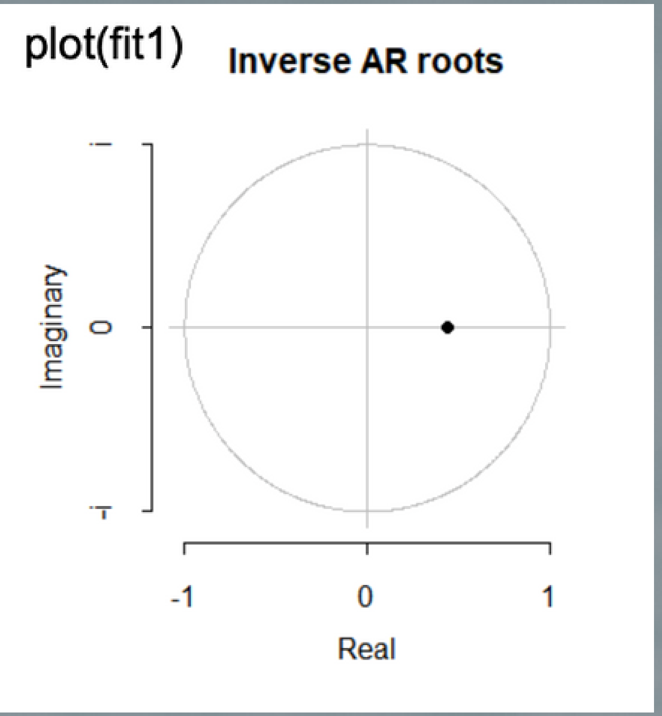
```
test1 <- Box.test(res1,fitdf=1,lag=2,type="Ljung-Box")
test2 <- Box.test(res2,fitdf=2,lag=3,type="Ljung-Box")
test3 <- Box.test(res3,fitdf=3,lag=4,type="Ljung-Box")
test4 <- Box.test(res4,fitdf=2,lag=3,type="Ljung-Box")
test5 <- Box.test(res5,fitdf=3,lag=4,type="Ljung-Box")
test6 <- Box.test(res6,fitdf=4,lag=5,type="Ljung-Box")
test7 <- Box.test(res7,fitdf=1,lag=2,type="Ljung-Box")
test8 <- Box.test(res8,fitdf=2,lag=3,type="Ljung-Box")

test1$p.value
test2$p.value
test3$p.value
test4$p.value
test5$p.value
test6$p.value
test7$p.value
test8$p.value
```

Test	p–Value
ARIMA (1,0,0)	0.1908183
ARIMA (1,0,1)	0.03964004
ARIMA (1,0,2)	0.0002010581
ARIMA (2,0,0)	0.0761668
ARIMA (2,0,2)	2.537756e-06
ARIMA (0,0,1)	0
ARIMA (0,0,2)	2.063974e-07

Setting up the null hypothesis which states that residuals are independent. It is observed that the p-value is significant for the first and fourth model i.e. AR(1) and AR(4). However, **as the value of the parameter increase, it leads to a fall in efficiency**. Hence, AR(1) model is chosen as the most appropriate

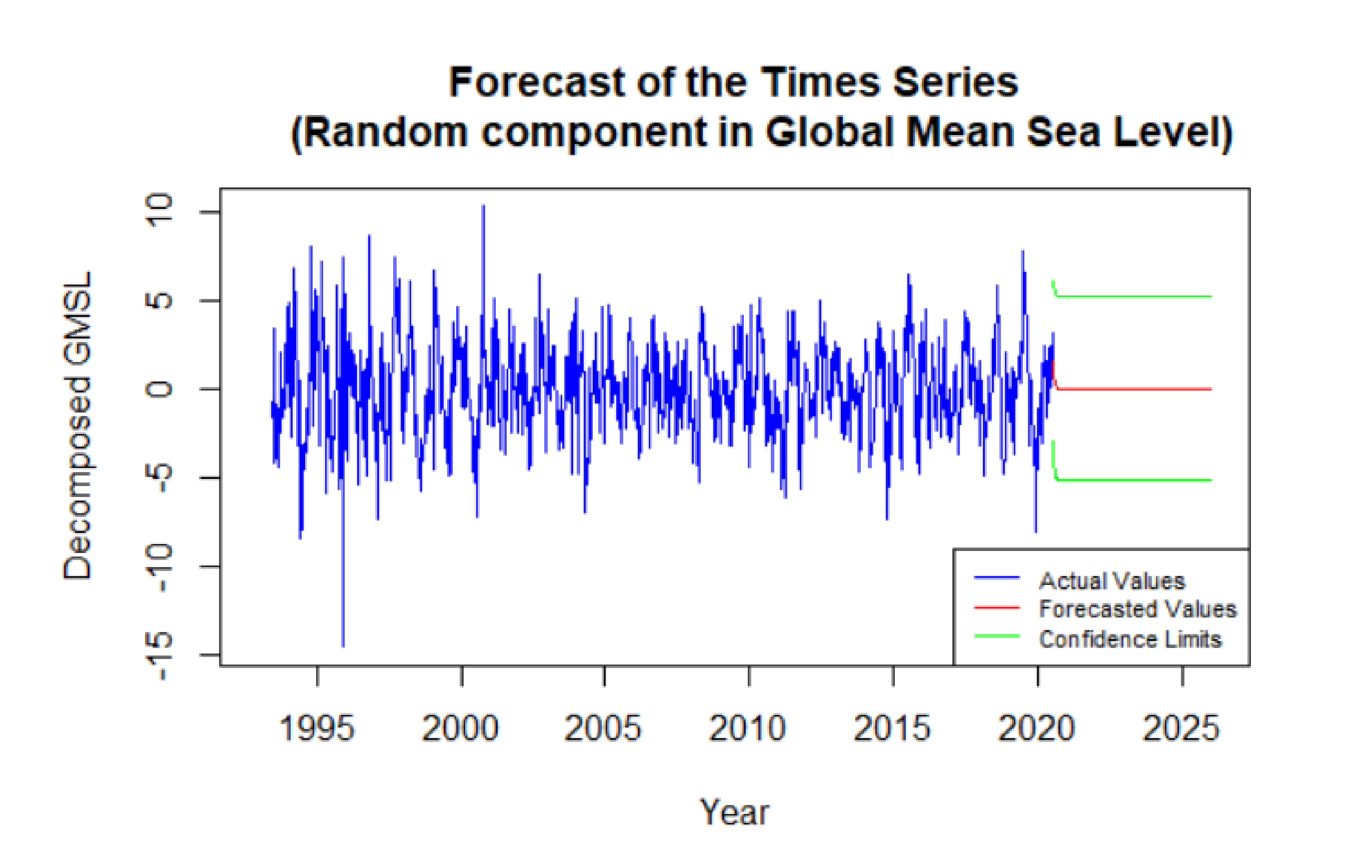
FITTED MODEL

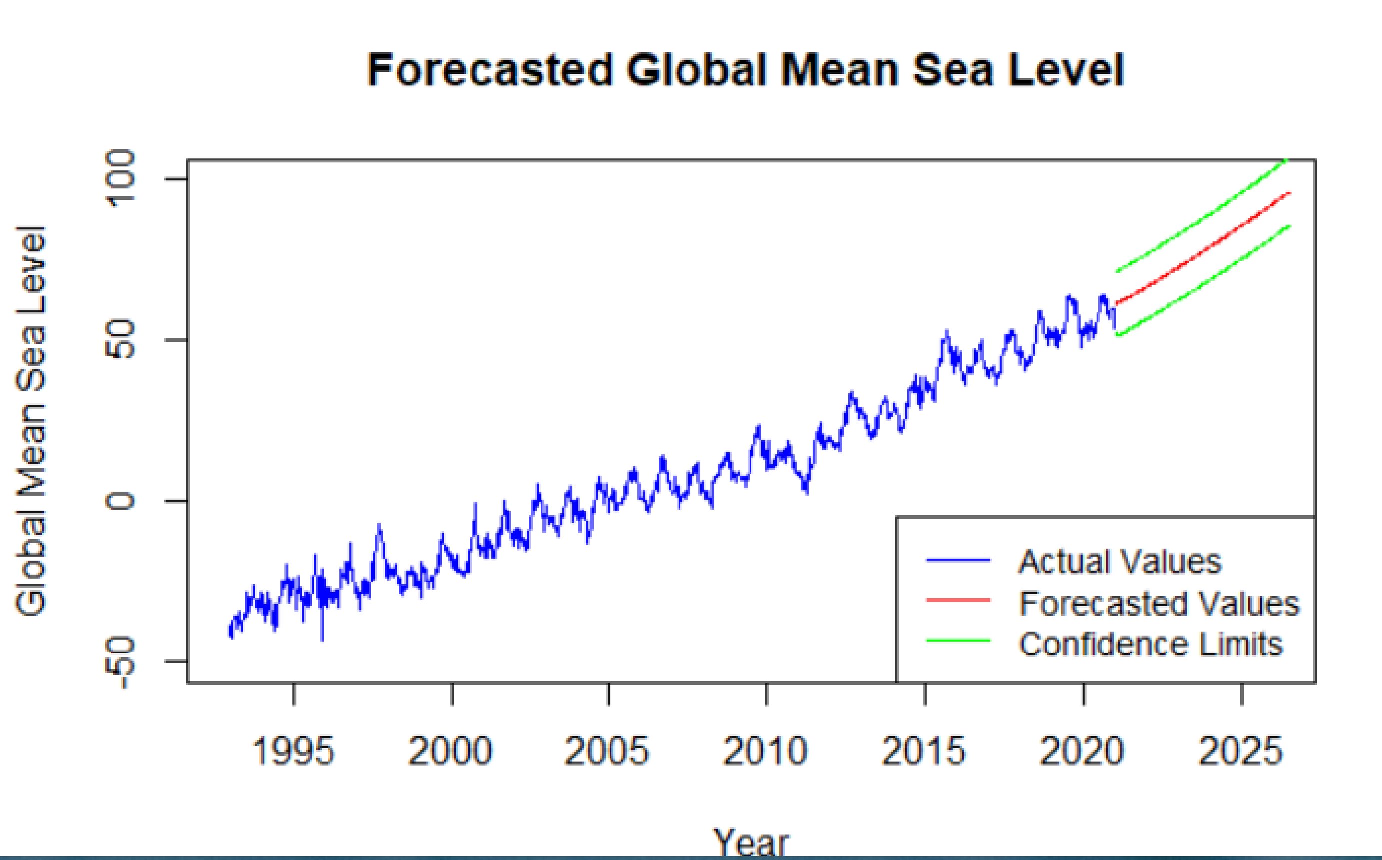
```
Call:
arima(x = ran.ts, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.5027    -0.0006
s.e.  0.0273     0.1451

sigma^2 estimated as 5.215:  log likelihood = -2244.84,  aic = 4495.68
>
```



It is known that for a series to be stationary, the modulus of unit roots should be less than or equal to 1. Since the roots lie within the unit circle, the series is stationary.

FORECASTING THE TIME SERIES AND PREDICTING THE FUTURE VALUES





The confidence limits have been calculated as:
(predicted value -1.96 x (standard error),
predicted value -1.96 x (standard error))

The forecasted values have been obtained by -

Forecasted Sea Level = Forecasted Trend Component + $\frac{\text{Forecasted Stationary Component}}{\text{Adjusted Seasonal Indices}}$

It can be observed that the sea level is further going to rise in the future.

LIMITATIONS OF THE STUDY

- The analysis is based on an additive model, for which it is assumed that all components are independent. However, this may not be the case in reality.
- There is a lack of data description from the source of data.

FUTURE SCOPE OF THE STUDY

- Performing detailed region-wise impact analyses to further understand and predict changes in sea levels in different water bodies across the world. This can support in designing prevention plans for such specific regions.
- Extending the study of changing sea-levels to better the existing understanding of natural disasters like tsunamis and floods.

INFERNCES FROM THE STUDY

Upon seeing the forecasted values, it is observed that the **sea level is further going to rise in the coming** future which may lead to more natural disasters such as tsunamis, posing as a threat to mankind and livelihood.

Hence, we as citizens should take active steps to ensure that sea level doesn't rise in the future. **This can be done by taking steps to reduce our carbon footprint and reducing our energy use.**

Individuals, organizations, and government leaders should work together to build **protections before flooding, build back stronger after flooding, and create plans that future-proof communities.**

REFERENCES

- The Analysis of Time Series by Chris Chatfield
- Fundamentals of Applied Statistics, S. C. Gupta and V. K. Kapoor
- Basic Econometrics by Damodar N. Gujrati, Dawn C. Porter and Sangeetha Gunasekhar
- Introduction to Time Series and Forecasting by Richard A. Davis
- Introducing Sea Level Change: *Introducing Earth and Environmental Sciences* by Alastair Dawson
- The Water Will Come: *Rising Seas, Sinking Cities, and the Remaking of the Civilized World* by Jeff Goodell
- https://www.kaggle.com/datasets/kkhandekar/global-sea-level-1993-2021
- https://www.earthdata.nasa.gov/resource-spotlight/sea-level-change
- https://www.isro.gov.in/InfoClimateEnv.html