1. Descriptive Analysis

There are 3 `-99.9`s (first, last, last but one) in the dataset, which shows there is no actual value recorded, so I replace them by NAs and summary the dataset.

Table 1: Summary of the Dataset

DJF	MAM	JJA	SON
Min. :-1.200	Min. : 5.600	Min. :13.10	Min. : 7.500
1st Qu.: 2.900	1st Qu.: 7.600	1st Qu.:14.80	1st Qu.: 9.200
Median : 3.800	Median : 8.200	Median :15.30	Median : 9.700
Mean : 3.747	Mean : 8.169	Mean :15.31	Mean : 9.709
3rd Qu.: 4.700	3rd Qu.: 8.800	3rd Qu.:15.88	3rd Qu.:10.300
Max. : 6.800	Max. :10.300	Max. :17.80	Max. :12.600
NA's :1	NA	NA's :1	NA's :1

We can see the general order in temperature `JJA (summer) > SON (autumn) > MAM (spring) > DJF (winter)`

2. Overall trend

Here I use two ways to detect the overall trend of the data: simple linear regression and kernel smoothing (`bandwidth = 40/n`) (10 years).

2.1 Linear Regression

The linear model we are going to fit is:

$$y_t = \alpha_0 t + \alpha_1 t^2 + \alpha_2 t^3 + \beta_1 Q_1(t) + \beta_2 Q_2(t) + \beta_3 Q_3(t) + \beta_4 Q_4(t) + \epsilon_t$$

In which the indicator functions $Q_1(t)$ to $Q_4(t)$ correspond with winter, spring, summer and autumn and `t` ranges from 1 to the last season recorded. And by testing that the t^4 term will not be significant, so I only take the highest as cubic term in the model. Here are the model results:

```
        Estimate Pr(>|t|)
t        0.002537  4.0e-05
I(t^2) -0.000004  1.8e-05
I(t^3)  0.000000  0.0e+00
Q1       3.154715  0.0e+00
Q2       7.577911  0.0e+00
Q3      14.722951  0.0e+00
Q4       9.118443  0.0e+00
```

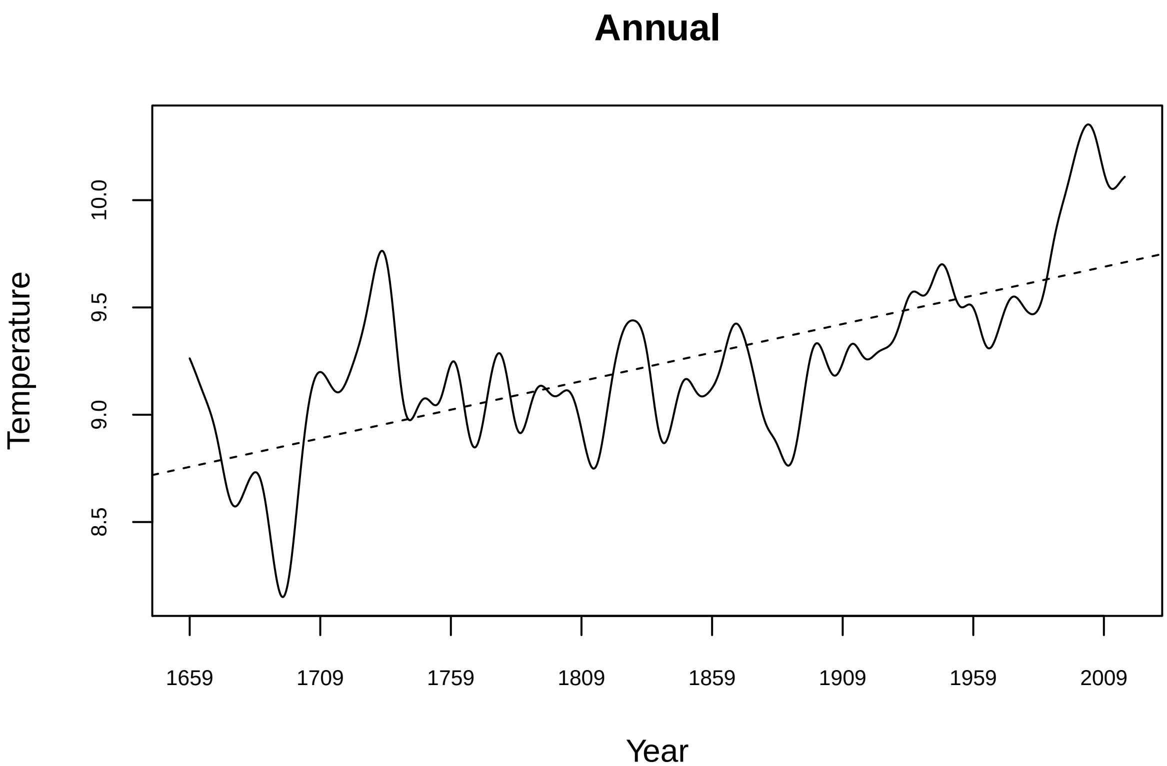
So the whole model is:

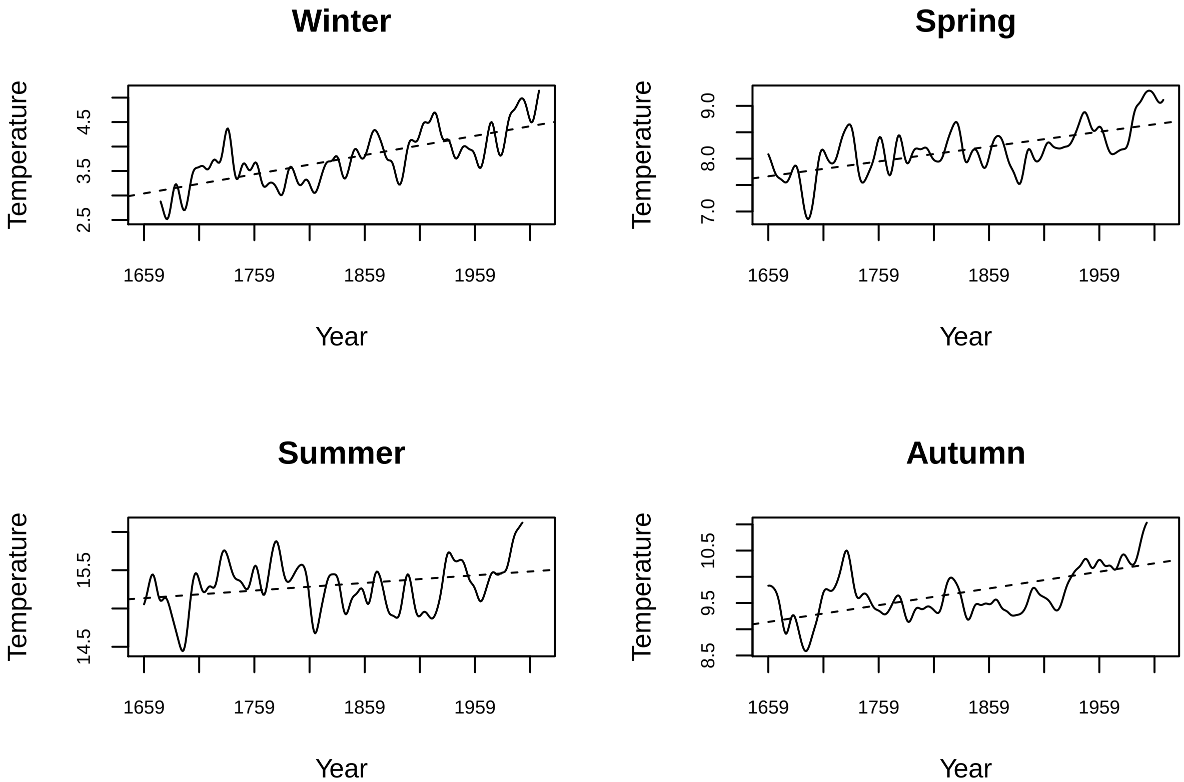
$$\hat{y}_t = 2.54 * 10^{-3}t - 4.29 * 10^{-6}t^2 + 2.32 * 10^{-9}t^3 + 3.15Q_1(t) + 7.58Q_2(t) + 14.72Q_3(t) + 9.12Q_4(t)$$

The whold model reaches R-Sqaured 0.99, so it fits pretty well. and the p-value for estimated coefficient of all `t` terms are extremely small, so there is significant increasing trend in the temperature by years.

2.2 Plots for annual and for different seasons

Combining linear models and kernel smooting with bandwidth 40/n (10 years), I make plots for annual temperature and seasonal temperature seperately:

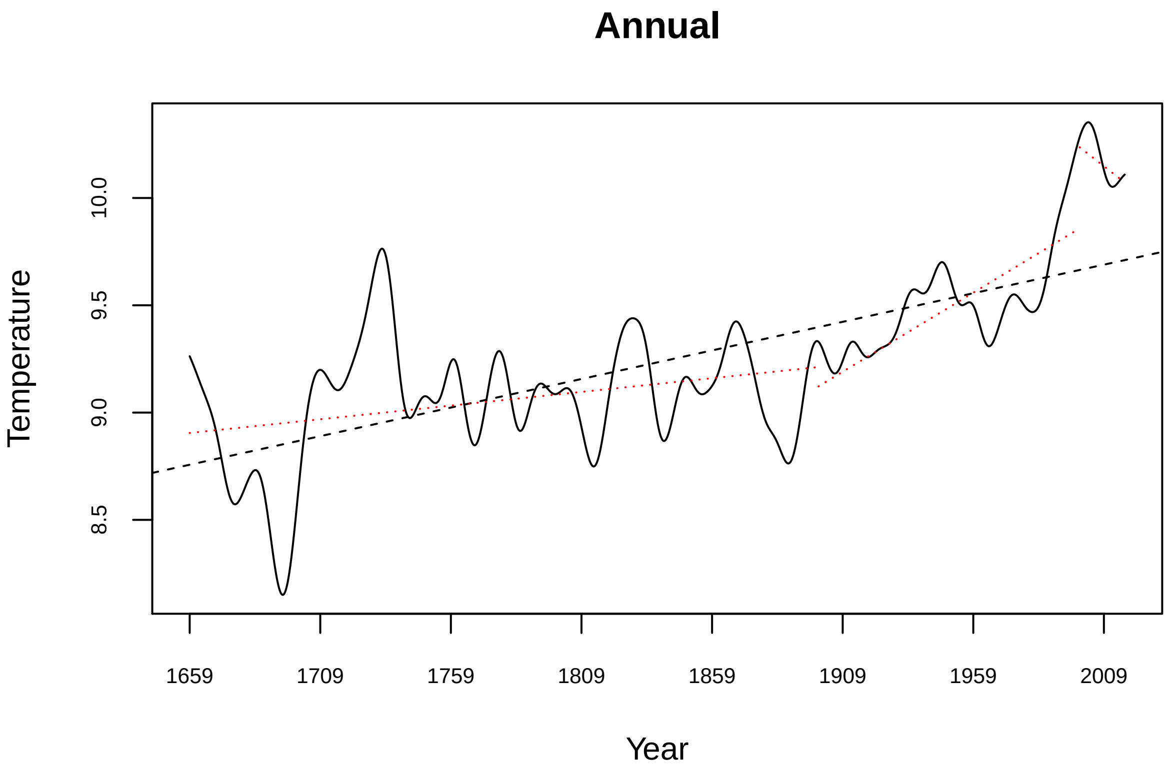




From both methods, we can see the increasing trend in temperature over years, though the increase in summer is the smallest, it is still significant under 5% level.

2.3 Did global warming slow down in 21st century?

In order to testify that claim, I divide the whole time period into three parts, 1659~1899, 1900~1999 and 2000~2017, and do three linear regression. The estimated lines are shown as the red dotted lines.

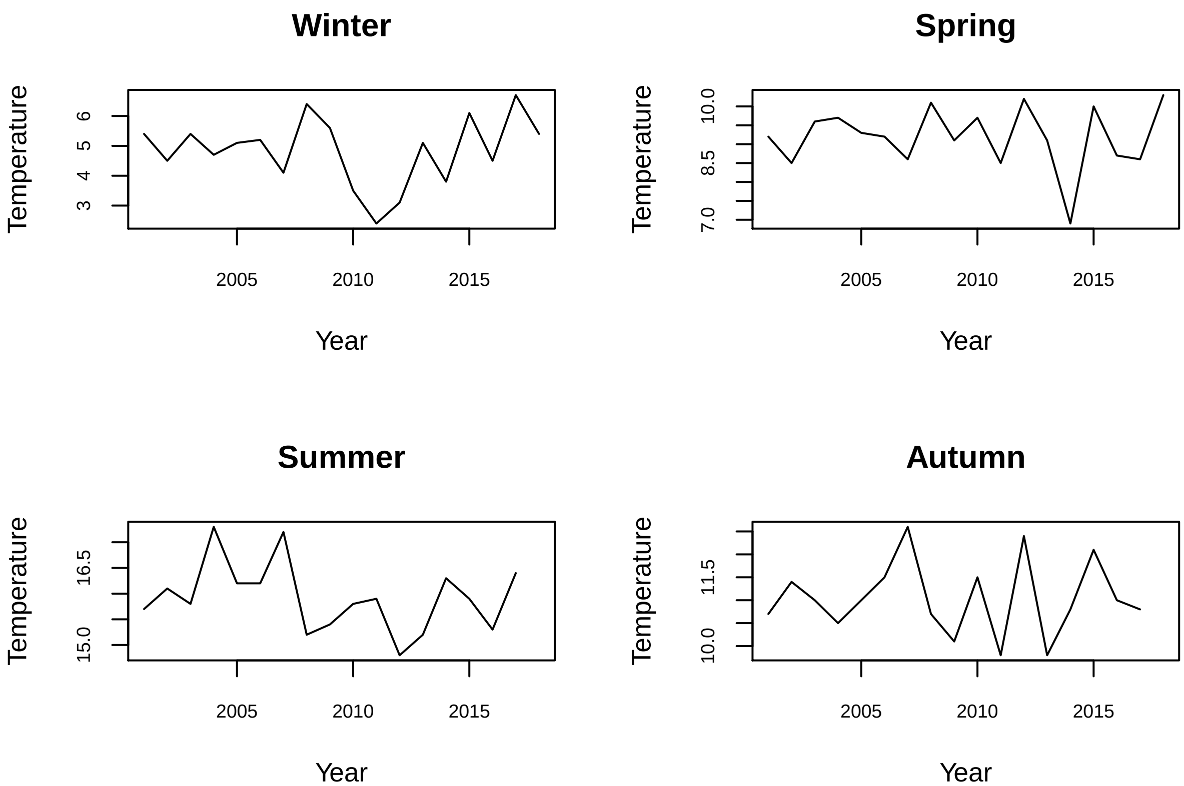


We can see that the overall trend in 21st century is decreasing rather than increasing, so here comes the question, in which season the temperature decreases the most?

The coefficients and plots for different seansons in 21st century are in Table2 ad following figure:

Table 2: Coefficients for different seasons

winter	spring	summer	autumn
0.005366	-0.006914	-0.03578	-0.001961



We can see that the main cause of decrease in 21st is in summer, which the coefficient is -0.036, much larger than other seasons in absolute value.

However, using only the simplest linear regression and with 17 years' data may not be persuasive, which we can see from the plot that the only reason for summer has such large negative coefficient is that there two relatively high values. More importantly, the time predictor in all 5 linear models above have very large p-values, so we cannot confirm the decreasing trend of temperature in 21st century.

Although the temperature trend in 21st century may not be convincing enough to reverse the global warming, we can still see that comparing to what has happened in 20th century, the pace has significantly slowed.

3. Detrending

3.1 Linear Regression Detrending

Thus, before further analysis, we need to detrend the data. We have two ways of detrending the series to make it stationary. One is using the result of linear regression by substracting the `t` term in original series:

$$y_t - \alpha_0 t - \alpha_1 t^2 - \alpha_2 t^3 = \beta_1 Q_1(t) + \beta_2 Q_2(t) + \beta_3 Q_3(t) + \beta_4 Q_4(t) + \epsilon_t$$

After detrending the data in this method, we again do the linear regression with `t`:

	Estimate	Pr(>|t|)
t	0.00	0.96
Q1	3.26	0.00
Q2	7.68	0.00
Q3	14.82	0.00
Q4	9.22	0.00

We can clearly see that now the series is not dependent on `t`.

3.2 Differencing

Lag=1

Another way of detrending is to use differencing method. And in the seasonal model, we can either differencing with `lag=1` or differncing with `lag=4`. Here are the results of linear regression for `lag=1`:

$$y_t - y_{t-1} = \gamma_1 dQ_1(t) + \gamma_2 dQ_2(t) + \gamma_3 dQ_3(t) + \gamma_4 dQ_4(t) + \epsilon_t$$

	Estimate	Pr(>|t|)
t1	0.00	0.94
dQ1	4.42	0.00
dQ2	7.15	0.00
dQ3	-5.61	0.00
dQ4	-5.97	0.00

Here it should be explained that because of the differencing method, we can no longer use the notation of every season, instead, here `dQ1` means winter to spring, `dQ2` spring to summer, `dQ3` summer to autumn and `dQ4` autumn to winter. The `t1` here, which stands for the time, is not significant, the R-Squared of the model is 0.96, so the model fits well.

Lag=4

When we use `lag=4`, the original seanson will lose their meaning, so in this model, I only regress the series with time `t`, here are the results:

```
            Estimate Pr(>|t|)
(Intercept) -0.00104  0.98805
t2           0.00001  0.90650
```

It is obvious that after differencing for both methods, the data also gets rid of its trend.

4. Seasonal ARIMA

4.1 General Model

To connect with what we have learnt from the class, I apply the differencing method in the further analysis. In this case, we have differencing and seasonal traits, so we first look at the general situation $ARIMA(p, d, q)(P, 1, Q)_4$ ($d = 0$ *or* 1), since one year has four seasons, it is reasonable to set number of periods per "season" (namely a year) as 4.

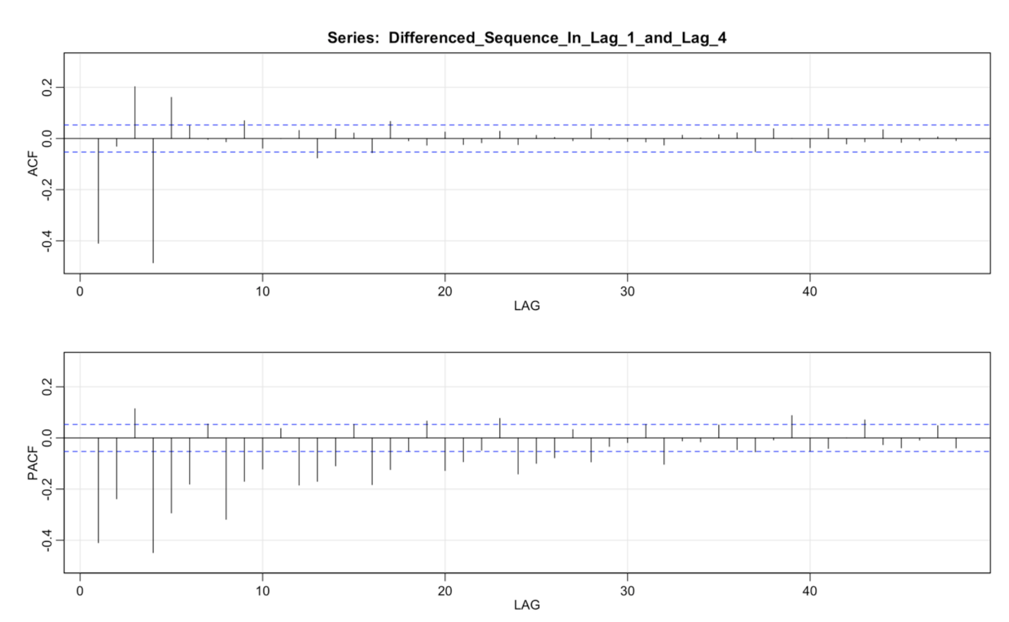
The model is as followed:

$$\phi(B)\Phi(B)(1-B)^d(1-B^4)x_t = \theta(B)\Theta(B^4)w_t$$

(From left to right: Non-seasonal AR(p), SAR(P), Non-seasonal differnece(d), Seasonal difference(1), MA(q), SMA(Q))

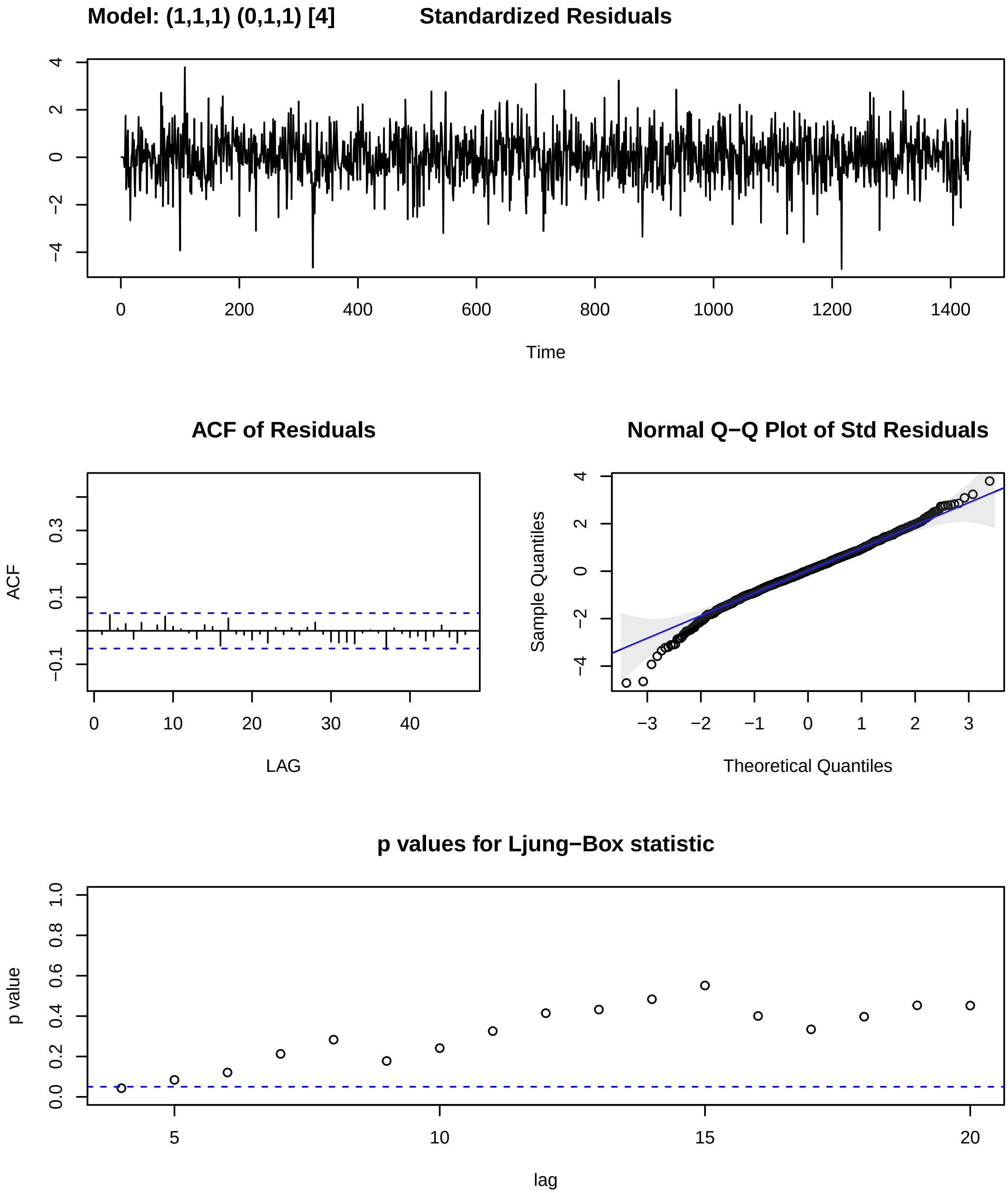
4.2 `sarima(1,1,1,0,1,1,4)`

We look at the acf and pacf plots after detrending and taking seasonal difference:



First consider about the seasonal lags, it is obvious that in seasonal lags, ACF cuts off after `lag=4`, so $Q = 1$, and PACF tails off at seasonal lags, so $P = 0$. Then, consider about the within seasonal lags. Both ACF and PACF for `lag=1` are significant, we can try models ARMA(1,1), ARMA(2,1) or ARMA(2,3). In order to simplify the model, we should try ARMA(1,1) first to see whether it is enough for fitting the data. Based on the following outputs, we can see that the ARMA(1,1) model fits good, so there is no need to consider more complicated model.

The following are the plots for residual analysis for the model along with their estimated coefficients, log-likelihood and AIC.

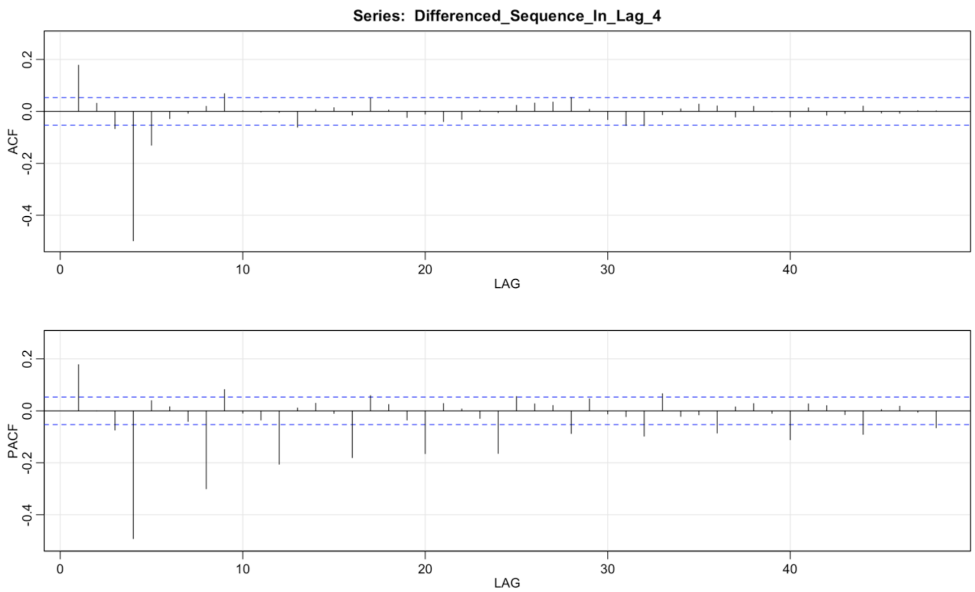


```
     Estimate     SE   t.value p.value
ar1    0.1770 0.0295    6.0041       0
ma1   -0.9631 0.0123  -78.6139       0
sma1  -0.9810 0.0061 -161.8443       0
```

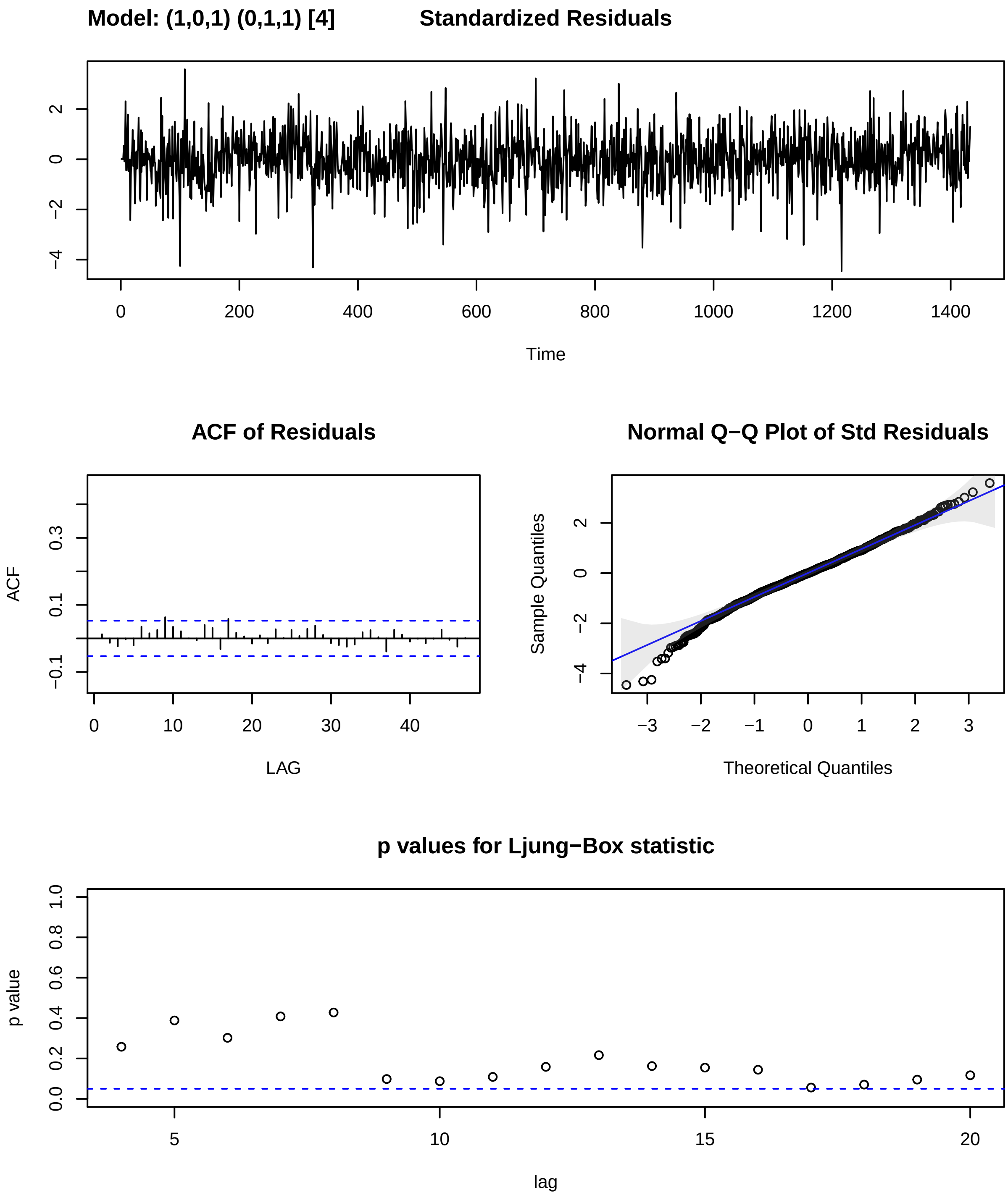
with $\hat{\sigma}^2 = 0.8827$, $loglikelihood = -1946.71$ and $AIC = 0.8794$.

4.3 `sarima(1,0,1,0,1,1,4)`

Similarly, we can do the same method with serie only taken seasonal differencing. Here is the ACF and PACF plot:



Same as 4.2, we can see that in seasonal lags, ACF cuts off after `lag=4`, so $Q = 1$, and PACF tails off at seasonal lags, so $P = 0$, and in this case, within seasonal lags, we can be certain about the ARMA(1,1). So the following are the plots for residual analysis for the model along with their estimated coefficients, log-likelihood and AIC for `sarima(1,0,1,0,1,1,4)`.



	Estimate	SE	t.value	p.value
ar1	0.6773	0.1051	6.4467	0.0000
ma1	-0.4891	0.1259	-3.8839	0.0001
sma1	-0.9696	0.0086	-112.4561	0.0000
constant	0.0009	0.0003	2.5963	0.0095

with $\hat{\sigma}^2 = 0.8787$, $loglikelihood = -1940.79$, $AIC = 0.8777$.

Both models' results are similar, nearly all lags of autocorrelation are close to 0, and the p-value are larger than 0.05, meaning that we cannot reject the null hypothese that the autocorrelation functions between residual are 0, there are few outliers in the normal Q-Q plot and standardized residuals plot, except these, both model fit well, .

5. Forecast

The model formula for `sarima(1,1,1,0,1,1,4)` is:

$$(1-0.177B)(1-B)(1-B^4)temp_t = (1-0.96B)(1-0.98B^4)w_t$$

And the formula for `sarima(1,0,1,0,1,1,4)` is

$$(1-0.677B)(1-B^4)temp_t = (1-0.489B)(1-0.97B^4)w_t$$

And the next 14 seasons predictions (from 2017 Summer to 2020 Autumn) along with the plots are (left is sarima(1,1,1,0,1,1), right is sarima(1,0,1,0,1,1)):

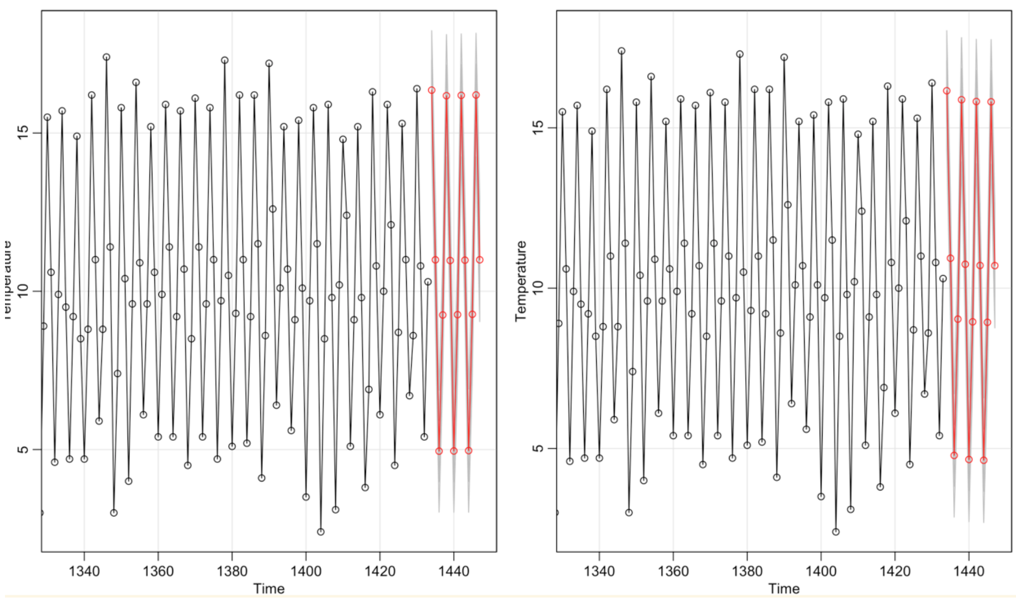


Table 3: Forecast with `sarima(1,1,1,0,1,1)`

	Winter	Spring	Summer	Autumn
2017	5.4	10.3	16.35	10.99
2018	4.95	9.252	16.18	10.97
2019	4.955	9.262	16.19	10.98
2020	4.967	9.273	16.2	10.99

Table 4: Forecast with `sarima(1,0,1,0,1,1)`

	Winter	Spring	Summer	Autumn
2017	5.4	10.3	16.16	10.93
2018	4.783	9.035	15.88	10.75
2019	4.657	8.95	15.82	10.71
2020	4.633	8.935	15.81	10.7

From the predictions we can see that, the temperature predicted in `sarima(1,1,1,0,1,1)` are higher than in `sarima(1,0,1,0,1,1)`, meanwhile, both models suggest the temperature to be mildly decreasing. Futhermore, we can see that the change in following years gradually becomes negligible, which is basically due to prediction of any time series will probably converge to constant.