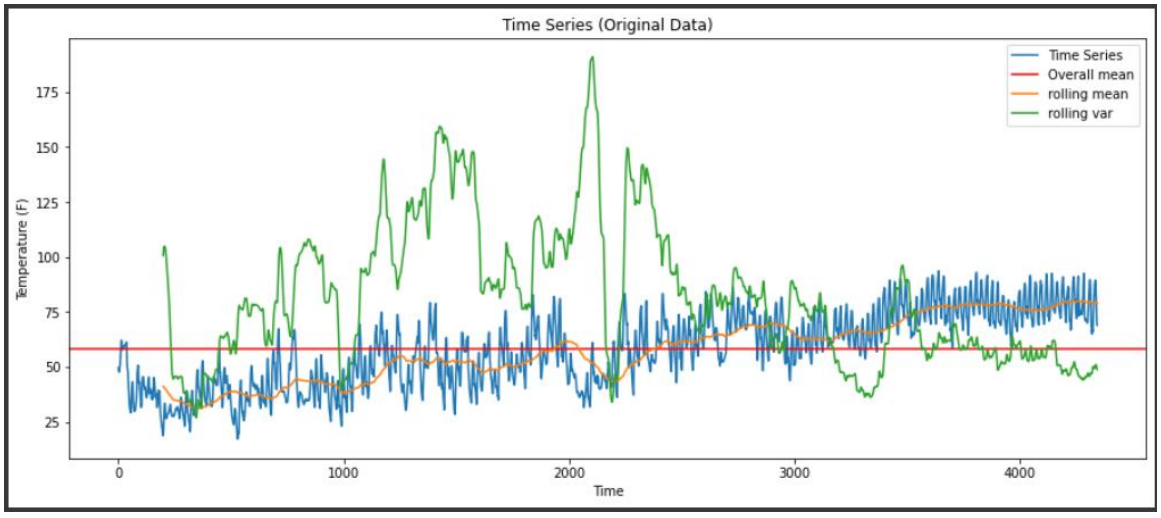
FORECASTING RESULTS

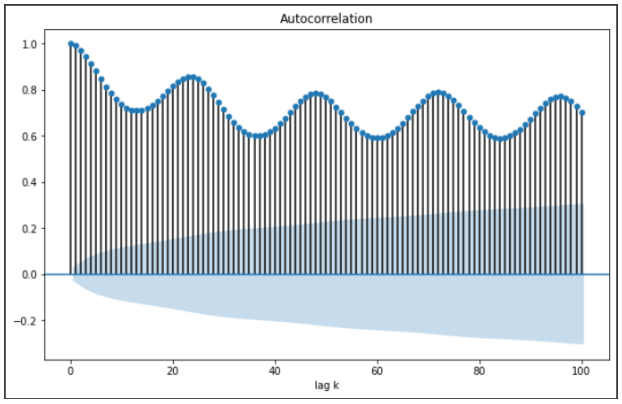
File used – 2.csv

TASK1: Check for stationary

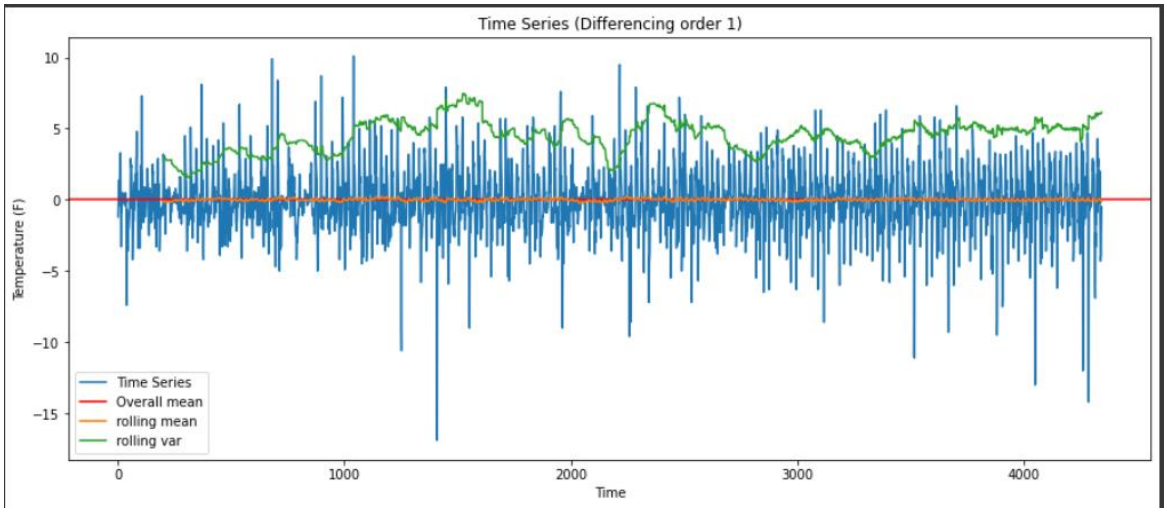
Original data:



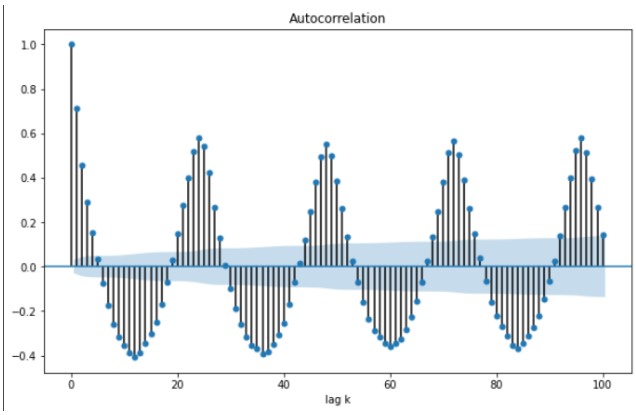
From the time series it is evident, that the original data shows a trend. The rolling mean of 200 values is no where near the overall mean. The variance is also not constant. There appears to be some seasonality, but we aren't sure of that. Thus, we plot an ACF plot for that. From the ACF it is evident that seasonality exists in this data (T= 24 hour). Data is not stationary.



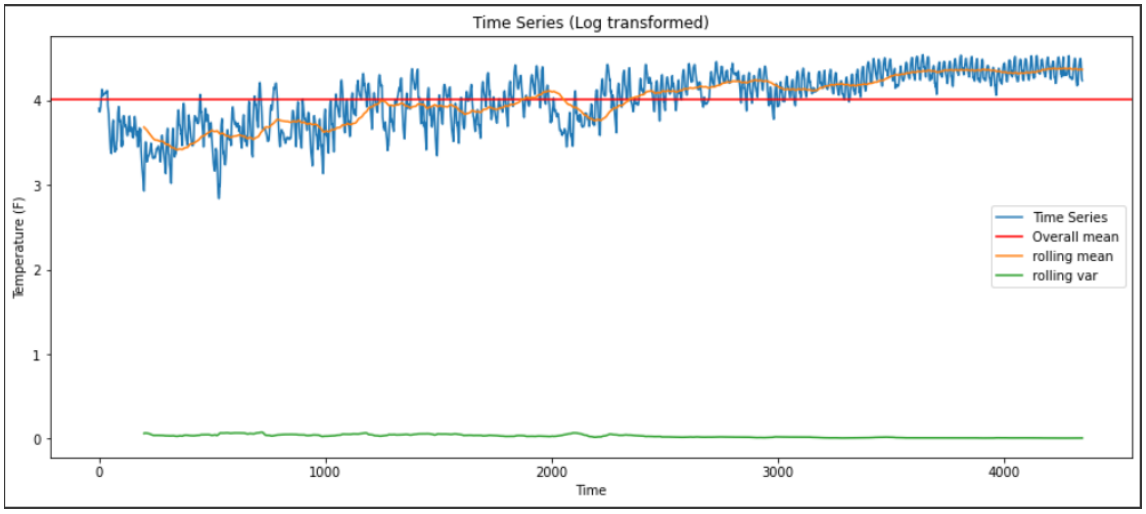
Differencing order 1:



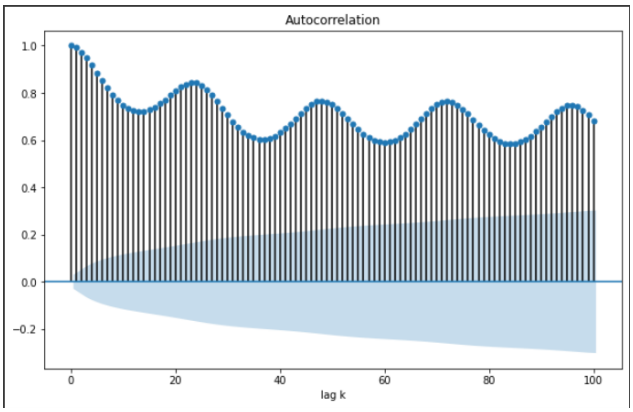
Differencing order 1 removes the trend. However. The variation is still not constant and from the ACF plot, we can also observe that seasonality still exists. Thus, this data is still not stationary.



Log transformation:

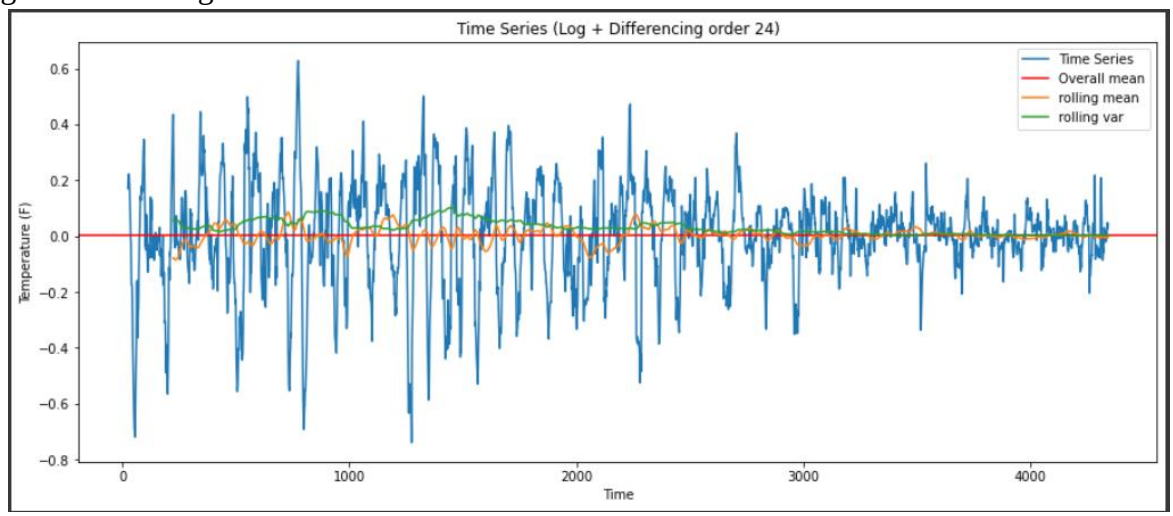


The log transformation removes the variation trend form the data. However, the trend (mean) still exists and ACF shows that seasonality still there. The data is still not stationary.

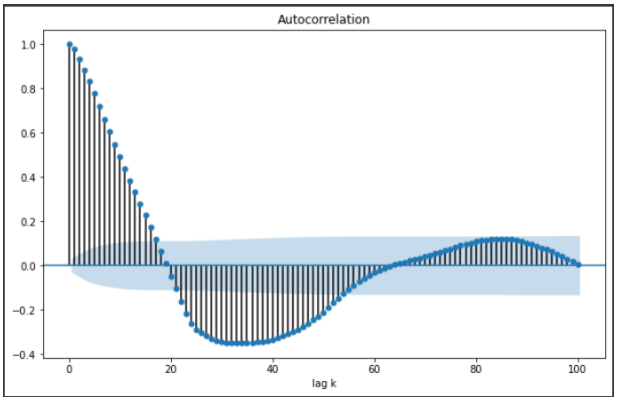


We will try to remove the trend, variation, and seasonality by first doing a log transformation, then a differencing order 24 (Time period of seasonality).

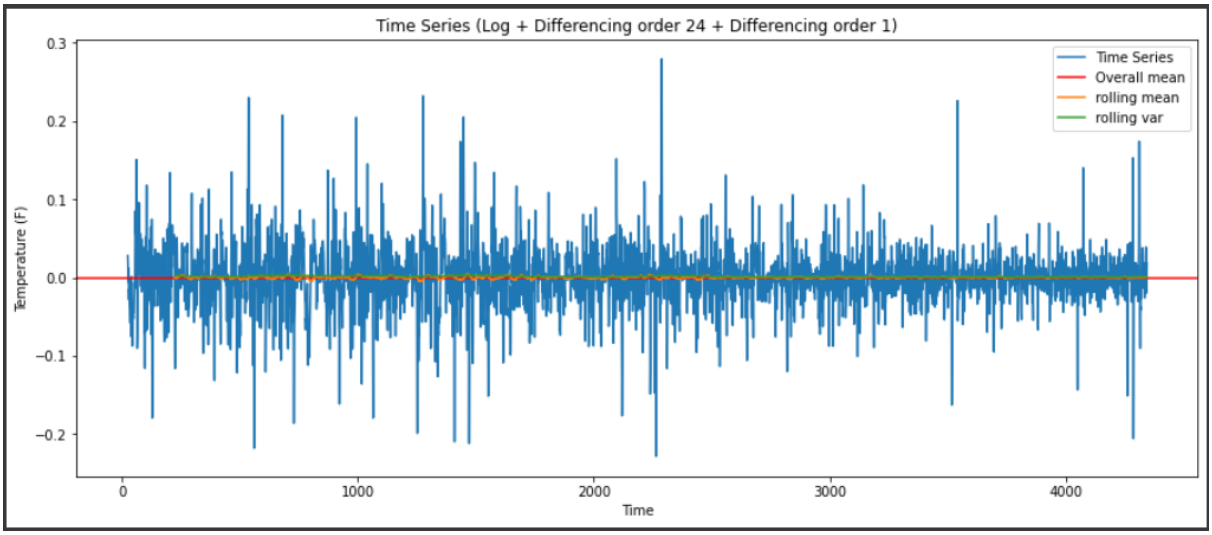
Log + Differencing order 24 transformation:



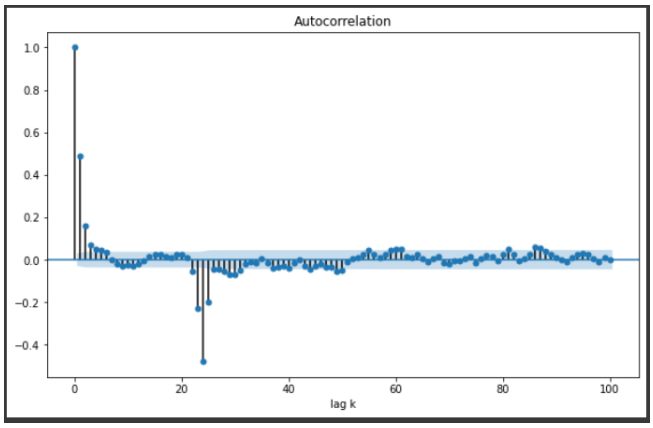
The trend still exists but the extent of it is much less than the original data. The same holds true for the variation. The ACF shows that although the seasonality still exists, the magnitude of peaks have started to die down as k is increases. Thus, the data is still not stationary.



Log + Differencing order 24 + Differencing order 1transformation:

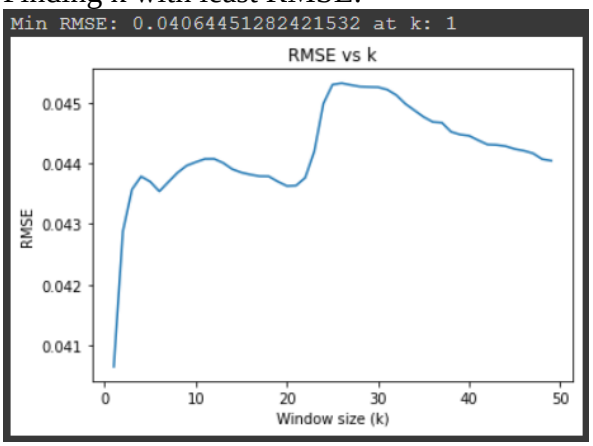


The time series shows that the trend and variation in the data is completely removed. The data looks like random noise, which is desirable. From the ACF we can see that there is no seasonality in the data as well. This data is, therefore, stationary.
We will use this data going forward for training and testing.

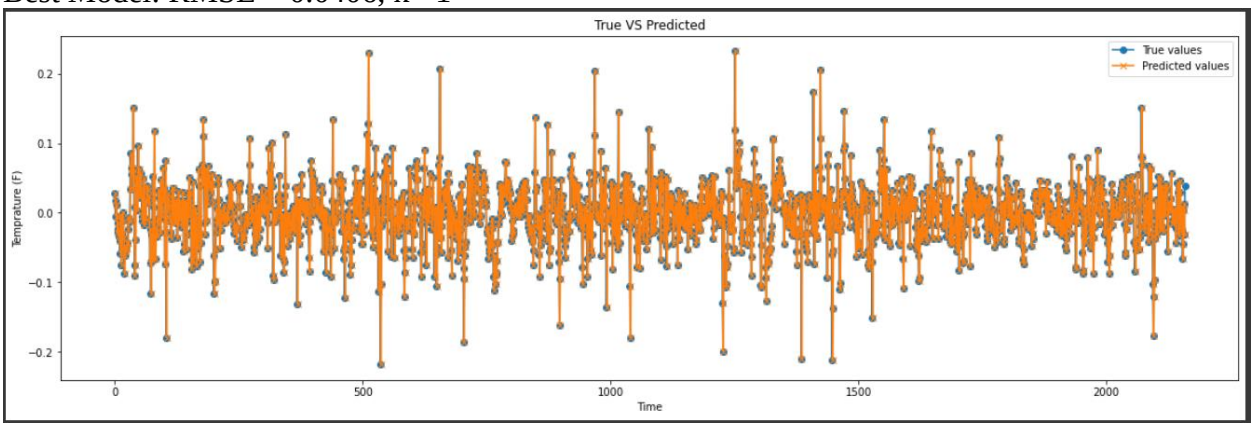


TASK 2: Simple moving average (train set)

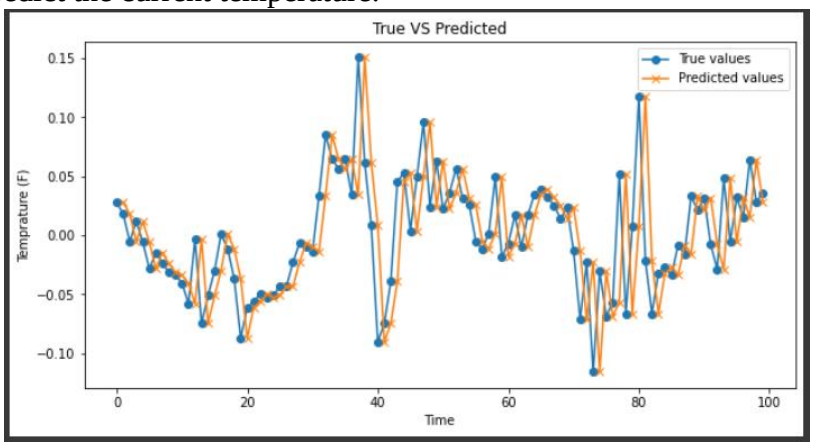
Finding k with least RMSE:



Best Model: RMSE = 0.0406, k =1



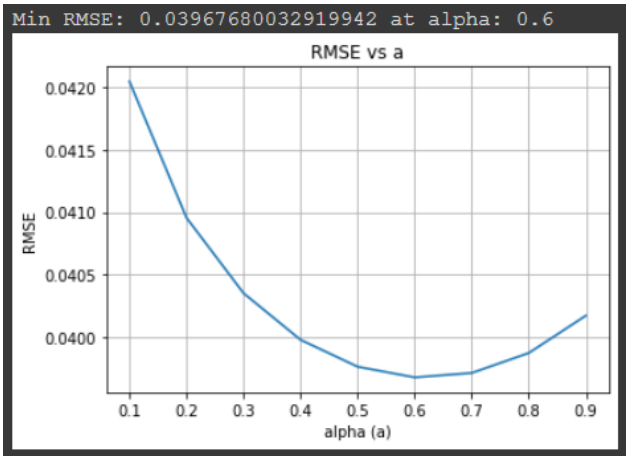
From the graph we can concur that the predicted values seem to follow the original values very closely. The best model comes out for k = 1, which means that the model just uses the previous temperature to predict the current temperature.



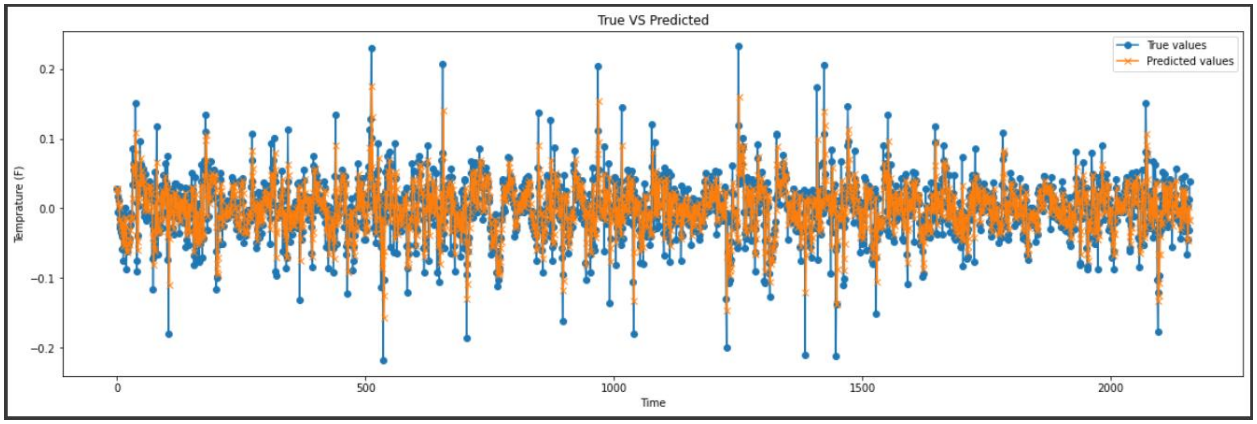
Looking at the first 100 values, we can see closely that the predicted curve is the same as the original curve with lag 1, which is understandable for k = 1.

TASK 3: Exponential smoothing model (train set)

Finding alpha with least RMSE:



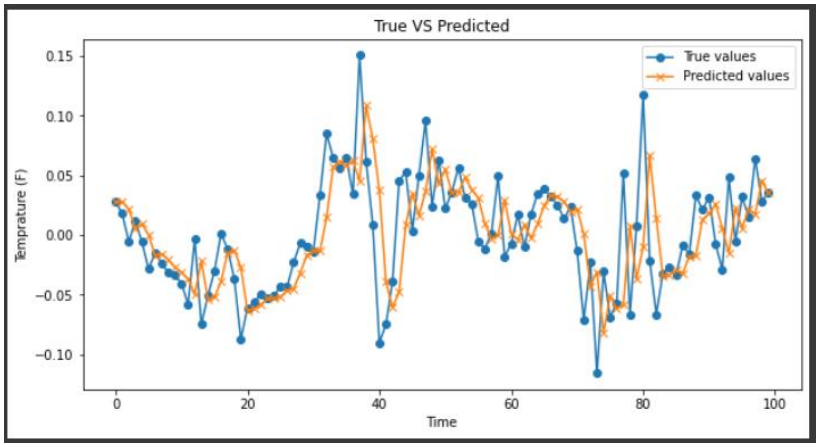
Best Model: RMSE = 0.03967, alpha = 0.6



From the graph we can see that the predicted model is unable to predict the extreme ends of the original values. This is more clearly visible from the magnified plot below. However, the RMSE for exponential model, is better than simple moving average.

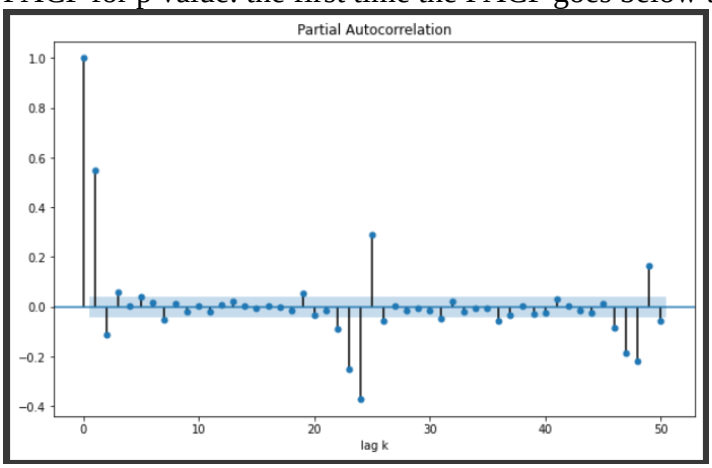
Our model gives the least RMSE for alpha=0.6, so our model equation becomes,

$s_t = 0.6*x_{t-1} + 0.4*s_{t-1}$. Thus, it is relying on both the previous original data, and the previous predicted value.



TASK 4: Autoregression model (train set)

PACF for p value: the first time the PACF goes below the confidence interval is for value p = 4



AR(4) summary:

AutoReg Model Results

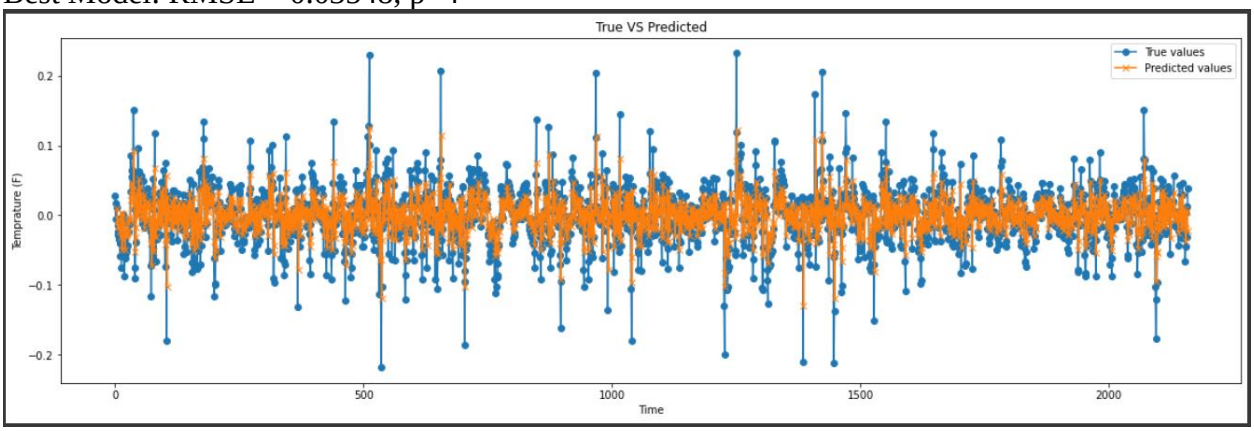
Dep. Variable:	Log-Diff24-Diff1	No. Observations:	2160
Model:	AutoReg(4)	Log Likelihood	4138.625
Method:	Conditional MLE	S.D. of innovations	0.035
Date:	Wed, 21 Oct 2020	AIC	-6.671
Time:	22:11:22	BIC	-6.656
Sample:	4	HQIC	-6.666
	2160		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.566e-05	0.001	-0.060	0.952	-0.002	0.001
Log-Diff24-Diff1.L1	0.6183	0.022	28.693	0.000	0.576	0.661
Log-Diff24-Diff1.L2	-0.1510	0.025	-5.966	0.000	-0.201	-0.101
Log-Diff24-Diff1.L3	0.0587	0.025	2.319	0.020	0.009	0.108
Log-Diff24-Diff1.L4	0.0040	0.022	0.186	0.852	-0.038	0.046

Roots

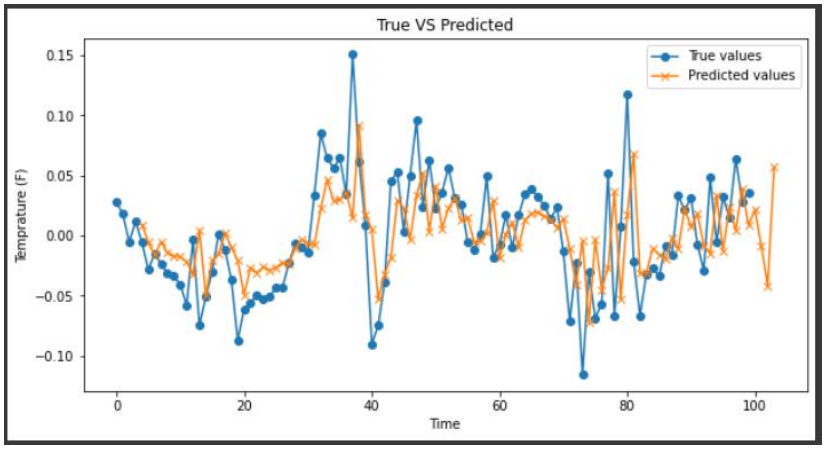
	Real	Imaginary	Modulus	Frequency
AR.1	1.7889	-0.0000j	1.7889	-0.0000
AR.2	0.4692	-2.7937j	2.8329	-0.2235
AR.3	0.4692	+2.7937j	2.8329	0.2235
AR.4	-17.3587	-0.0000j	17.3587	-0.5000

Best Model: RMSE = 0.03548, p=4



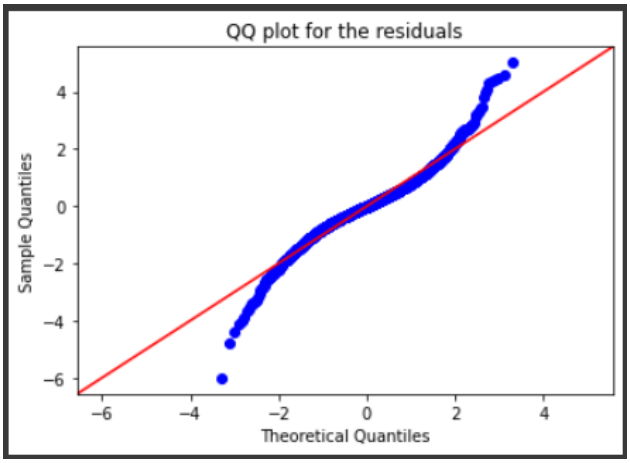
For AR(4) model, the equation of the model becomes, $X_t = -0.000045 + 0.6183*X_{t-1} - 0.1510*X_{t-2} + 0.0587*X_{t-3} + 0.0040*X_{t-4}$. From the p-values we can remove the const and X_4. Thus, the equation becomes, $X_t = 0.6183*X_{t-1} - 0.1510*X_{t-2} + 0.0587*X_{t-3}$.

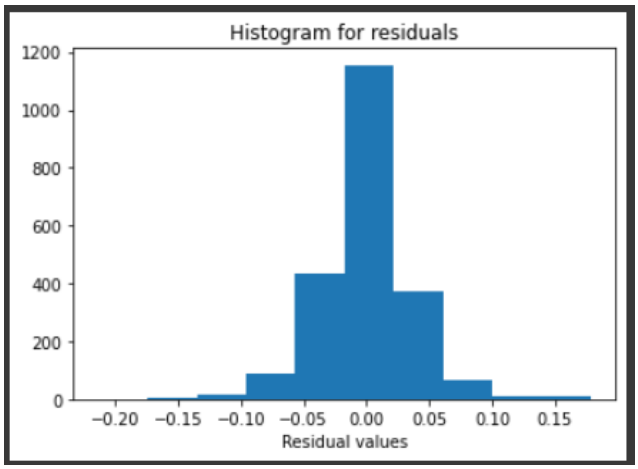
From the graph, we can see that similar to the exponential model, the predicted values from the AR model are not able predict the extreme ends of the original value. The AR model however has a lower RMSE than the exponential model, thus is performs better than both exponential and simple moving average models.



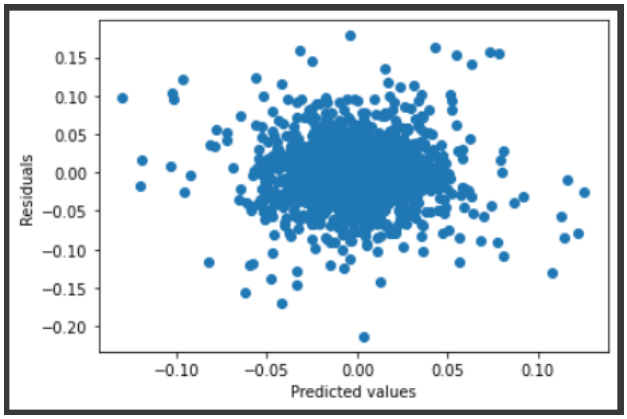
From the close-up it can be verified that the model is not able to correctly predict the extreme points of the original data.

Residual Analysis:





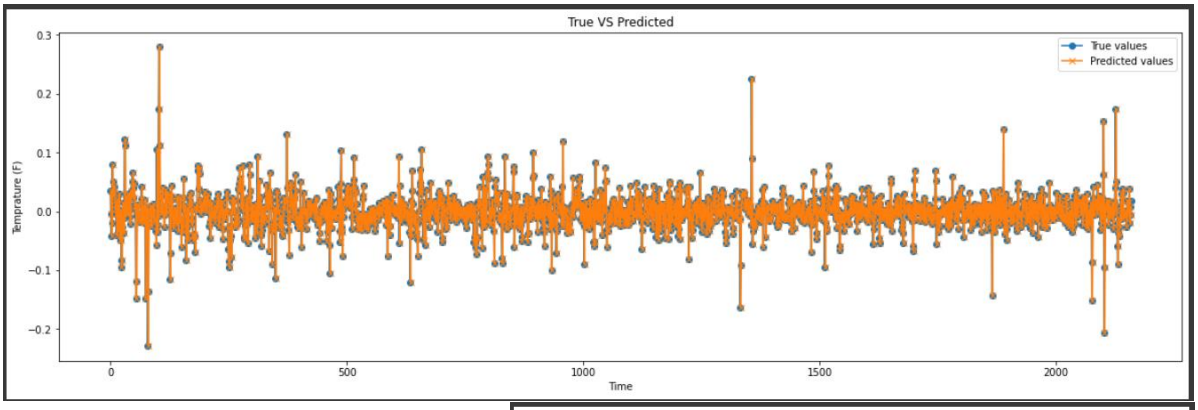

```
Chi-square test
pvalue: 9.68984396446292e-38
Null hypothesis rejected.
Residuals do not follow a normal distribution
```



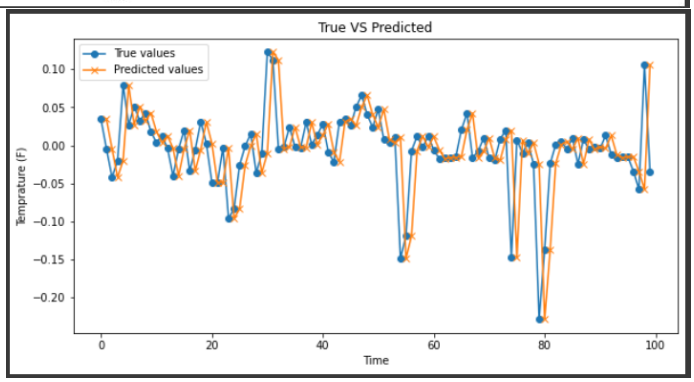
From the qq plot of the residuals, we can see that the residuals don't follow the 45-degree line through its entirety. Thus, it is unlikely that the residuals follow a normal distribution. This is further confirmed by the histogram plot of the residuals and the chi-square test.
The scatter plot of the residuals do not show any significant trend.

TASK 5: Comparison of all the models (test set)

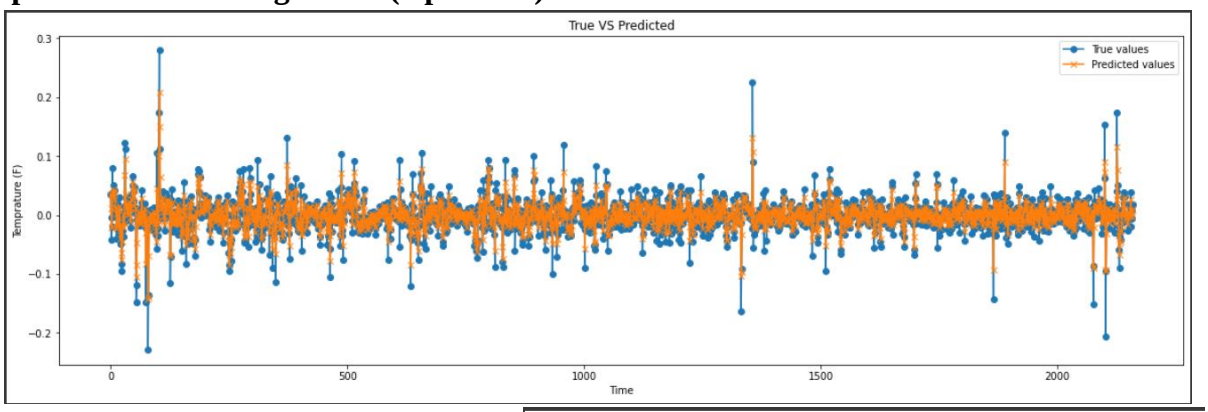
Simple moving average model (k=1): RMSE = 0.03376



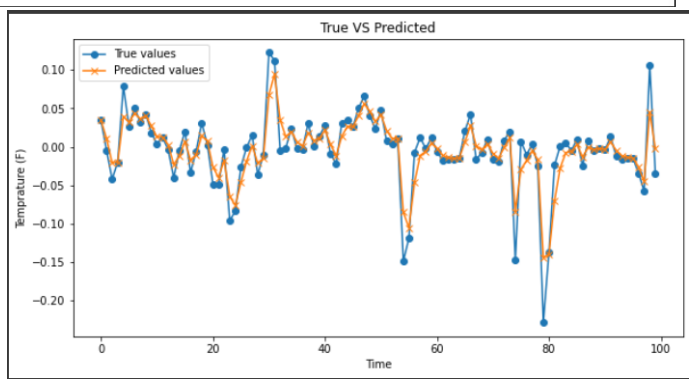
Similar to the training set, the test predictions also follow the original value by lag 1.



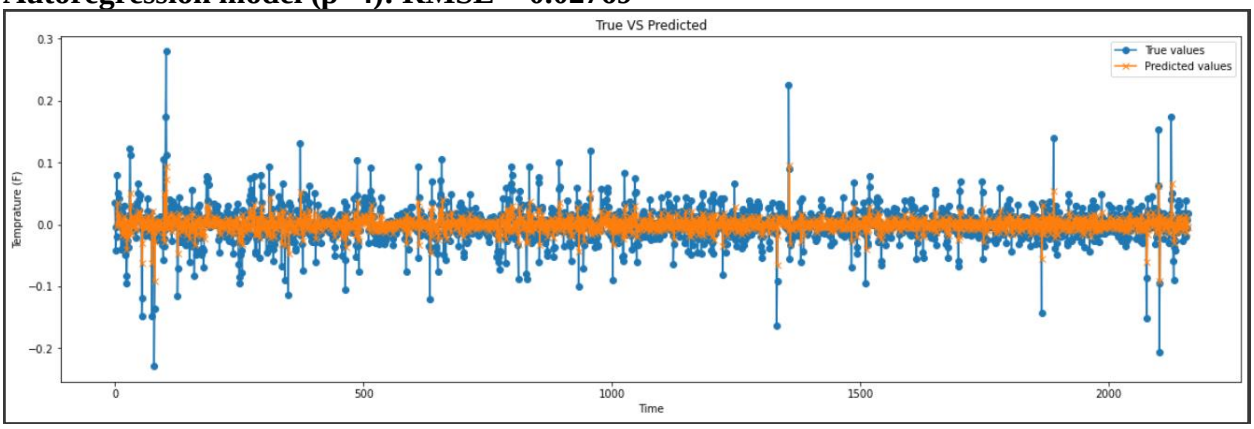
Exponential smoothing model (alpha=0.6): RMSE = 0.0126



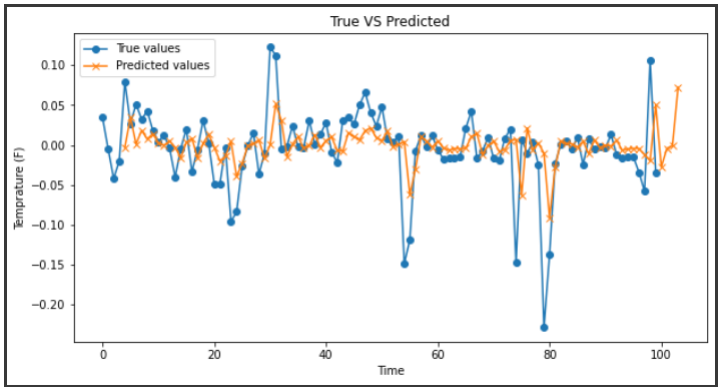
The exponential model predictions aren't able to predict the extreme points of the original data. It performs better than simple moving average.



Autoregression model (p=4): RMSE = 0.02769



The AR model, similar to exponential model isn't able to predict the extreme values of the original data. Also, the AR model performed better than exponential model for the training set. However, for the test set, it performs worse than the exponential model. It performs better than simple moving average though.



RMSE comparison:

Dataset	Simple moving average	Exponential smoothing	Autoregression
Train set	0.04064	0.0396	0.03548
Test set	0.03376	0.0126	0.02769

For the given data (after transformation), the simple moving average performs the worst for both the training and the testing sets. AR model performs the best on the training set, while exponential model is the best for the test set. Given the difference in the RMSE for these two models, for both train and test sets, we can conclude that overall, the exponential model yields the best predictions. The performance gain for AR model for train set (as compared to exponential model), is much less than the performance gain in the exponential model for test set (when compared to AR RMSE for test set).
The exponential mode has alpha=0.6, which means when predicting at a given time, more weightage is given to the original data at t-1 than the predicted value at t-1. This makes sense because we are dealing with hourly temperature and temperature at any time would be closely related to the previous original value.