Fall 2019

Seasonal Time Series Models with Application to Weather and Lake Level Data

Mengqing Qin
Missouri State University, Mengqing1214@live.missouristate.edu

SEASONAL TIME SERIES MODELS WITH APPLICATION TO WEATHER AND LAKE LEVEL DATA

A Master's Thesis

Presented to

The Graduate College of

Missouri State University

In partial Fulfillment

Of the Requirements for the Degree

Master of Science, Mathematics

By

Mengqing Qin

December 2019

SEASONAL TIME SERIES MODELS WITH APPLICATION TO WEATHER AND LAKE LEVEL DATA

Mathematics

December 2019

Master of Science

Mengqing Qin

ABSTRACT

This work studies seasonal time series models with application to lake level and weather data. The thesis includes related time series concepts, integrated autoregressive moving average models (abbreviated as ARIMA), parameter estimation, model diagnostics, and forecasting. The studied time series models are applied to the data of daily lake level in Beaver Lake (1988-2017) and the data of daily maximum temperature in New York Central Park (1870-2017). Due to seasonality of the data, three different approaches are proposed to the modeling: regression method, functional ARIMA method and multiplicative seasonal ARIMA method. The forecasted values of the year 2018 are compared with observations; regression method is better to forecast daily values, and multiplicative ARIMA method is a better choice owing to higher accuracy for a short term and shorter period.

KEYWORDS: : seasonal time series, AR model, MA model, ARMA model, ARIMA model, multiplicative seasonal ARIMA, forecast

SEASONAL TIME SERIES MODELS WITH APPLICATION TO WEATHER AND LAKE LEVEL DATA

By

Mengqing Qin

A Master's Thesis
Submitted to the Graduate College
Of Missouri State University
In Partial Fulfillment of the Requirements
For the Degree of Master of Science, Mathematics

December 2019

Approved:

Yingcai Su, Ph.D., Thesis Committee Chair

Shouchuan Hu, Ph.D., Committee Member

Songfeng Zheng, Ph.D., Committee Member

Julie Masterson, Ph.D., Dean of the Graduate College

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Yingcai Su, who always offers useful tips and advice whenever I'm confused and helps me to solve my problems. He always encourages me to improve models and tells me it's more important to be an expert in the field. Under his help, I am more practiced in time series and R. Besides, sincere thanks to Songfeng Zheng, who is willing to answer my questions even though he is not my advisor. In addition, I indebted to Mr. Michael Biggs and Ms. Maria Guerra (U.S. Army Corps of Engineers) for providing the Beaver lake water level data, and National centers for environmental information for granting the permission for using the weather data in New York Central Park. Finally, I would like to thank all my friends for their company and help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 General Concepts for Seasonality

Many scholars give empirical or theoretical concepts for seasonality in time series. In daily life, there are many time series data, especially in economic fields, where a phenomenon repeats after a regular period of time. William called it seasonal time series data in [5]. More precisely, for a time series Y_t, and Y_t is measured s times a year, $s \geq 1$, and s is called the period. For the observations per period s, there are probably distinct means and variances stated by Franses in [4]. For example, the monthly series of average temperature [6] in Dubuque, Iowa, is high for summers and low for winters, and the phenomenon repeats yearly, which gives a seasonal period of 12. Similarly, for my project, the daily lake level of Beaver Lake and the daily max temperature in New York Central Park repeat the respective phenomenon per year, giving seasonal time series with period of 365.

1.2 Definitions

Jonathan and Chan introduced the following definitions in [6]. For a time series $\{Y_t : t = 0, \pm1, \pm2, \pm3, \dots\}$, the mean is defined by

$$\mu_t = E(Y_t) \qquad \text{for } t = 0, \pm1, \pm2, \pm3, \dots$$

where μ_t is the expected value of the series at time t.
The variance function is defined by

$$Var(Y_t) = E[(Y_t - \mu_t)^2] \qquad \text{for } t = 0, \pm1, \pm2, \pm3, \dots$$

The autocovariance function, $\gamma_{t,s}$ is given by

$$\gamma_{t,s} = Cov(Y_t, Y_s) \qquad \text{for } t, s = 0, \pm 1, \pm 2, \pm 3, \ldots$$

where $Cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)]$, and when $t = s$, I have $\gamma_{t,t} = Var(Y_t)$
The autocorrelation function (ACF), $\rho_{t,s}$ is given by

$$\rho_{t,s} = Corr(Y_t, Y_s) \qquad \text{for } t, s = 0, \pm 1, \pm 2, \pm 3, \ldots$$

where

$$Corr(Y_t, Y_s) = \frac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t)Var(Y_s)}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}$$

Similarly, sample autocovariance function at lag k is defined as

$$\hat{\gamma_k} = \frac{1}{n}\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y}) \quad \text{for } k = 1, 2, \ldots$$

and sample autocorrelation function at lag k is defined as

$$\hat{\rho_k} = \frac{\hat{\gamma_k}}{\gamma_0} = \frac{\frac{1}{n}\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

$$= \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

ACF is one useful tool for examining the dependence between the current data and the past lag k data and further specify MA(q) models. If the autocorrelation function is zero after lag q, then q will be chosen as the order of MA model. However, for autoregressive model, the autocorrelations of AR(p) model will tail off after lag p instead of going to zero. Therefore, Partial autocorrelation function (PACF) at lag p is needed to specify AR(p)

model. PACF at lag k is defined by

$$\phi_{kk} = Corr(Y_t - \beta_1 Y_{t-1} - \beta_2 Y_{t-2} - \cdots - \beta_k Y_{t-k+1},$$

$$Y_{t-k} - \beta_1 Y_{t-k+1} - \beta_2 Y_{t-k+2} - \cdots - \beta_k Y_{t-1})$$

Table 1 summarizes the general behavior of ACF and PACF which are useful in specifying models.

Table 1: Behavior Summary of ACF and PACF for ARMA models

	AR(p)	MA(q)	ARMA(p,q),p>0,and q>0
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

However, for mixed ARMA models, ACF and PACF have infinitely many nonzero values, making it difficult to specify orders for ARMA models. Thus, the extended autocorrelation method is proposed, but it's not my main topic for this project.

1.3 Stationary

For a stochastic process $\{Y_t\}$, it is called weakly stationary if it satisfies the following two requirements:

1. $E(Y_t)$ is a constant function for any time t

2. $\gamma_{t,t-k} = \gamma_{0,k}$ for all time t and lag k

which also means $\gamma_{t,s}$ only depends on lag k instead of time t.

As we know, one very important example for stationary process is white noise, denoted as $\{e_t\}$. It is a sequence of independently and identically distributed random variables with mean zero and variance σ_e^2. It is a significant element for ARIMA models.

There are two methods to check stationarity for any time series. Obviously, according to the definition, the easiest one is to check its mean, and observe whether the variance change a lot for a long term. However, it only provides a basic judgement. The other one is called Augmented Dickey-Fuller(ADF) test which assumes the time series Y_t is non-stationary, and Y_t can be approximated by a stationary model after differencing. The null hypothesis of ADF test is non-stationary, and the alternative hypothesis is that Y_t is stationary. Running *adf.test()* in R, gives two critical values in ADF test statistics, p-value and lag k. If the p-value is ≤ 0.05, the null hypothesis is rejected, which means Y_t is stationary after lag k. Otherwise, it's non-stationary. Lag k means Y_t will be stationary on the past k lags from the first difference of the observations. For example, if $k = 1$, it means $Y_t - Y_{t-1}$ is stationary; if $k = 2$, it means $(Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$ is stationary. Note, all codes in R through this thesis are from [6, 9, 10, 11].

In order to better understand a full process of measures needed for a time series, I take out all lake level values of Jan 1st and June 1st of Beaver Lake from 1971 to 2017 as examples in Figure 1.

From Figure 1, the variances of values around 1990 are clearly different from 1970-1985 and years after 2000. Thus, data of the lake level on Jan 1st and June 1st is not stationary from observation. Actually, it's absolutely normal to have a time series being non-stationary. There are two main useful ways to transform non-stationary to be stationary. The first one is to take difference as mentioned previously. For example, Y_t is a non-stationary time series, a new time series W_t is obtained by taking d^{th} difference, where $W_t = \bigtriangledown^d Y_t$, and

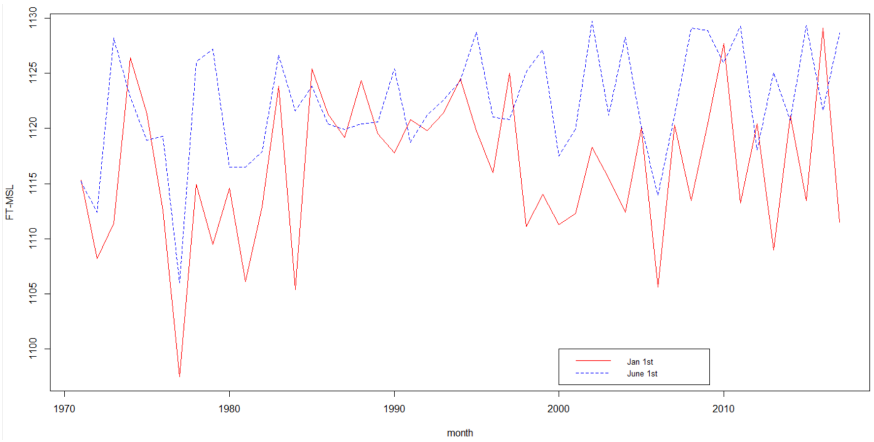

Figure 1: Beaver lake for Jan 1st and June 1st

W_t is stationary. When $d = 1$, $W_t = \bigtriangledown Y_t = Y_t - Y_{t-1}$. The second method is to take power transformation, introduced by Box and Cox in [3]. For any given parameter λ, the transformation is defined by

$$g(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log x, & \text{for } \lambda = 0 \end{cases}$$

The power transformation works only for positive data values. If there are some non-positive values, a positive constant must be added first to all of the data values to make them positive before taking power transformation. Based on power transformation, there is a special case proposed, taking logarithm, which is mostly used for percentage data. Besides, after logarithm is applied, the distribution of Y_t will behave better, and the effects of extrema and outliers will be reduced.

Based on Table 2, with Augmented Dickey-Fuller Test, after taking $Z_t = Y_t - Y_{t-1}$, the new time series becomes stationary with p-value$= 0.01 < 0.05$. Thus, I reject $H_0 : Z$ is non-stationary.

Table 2: ADF Test Results for Y and Z on Jan 1st

data	Y	Z
Dickey-Fuller	-2.7591	-6.0423
Lag order	3	3
p-value	0.2716	0.01
alternative hypothesis:	stationary	stationary

1.4 ARMA and ARIMA

Moving Average—MA

Suppose there is a time series Y_t, it could be written as the form of

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} = \sum_{j=0}^{q} \theta_j e_{t-j}$$

where $\theta_0 = 1$ and e_t is called white noise with mean zero and variance σ_e^2, and Y_t is called a moving average model with order q. Note, in R, you will get $-\theta e_1, \ldots, -\theta_q$ as the output.

According to section 1.2, for MA(1) model $Y_t = e_t - \theta e_{t-1}$,

$$E(Y_t) = E(e_t - \theta e_{t-1}) = 0$$

since e_t, e_{t-1} are white noise with mean zero. Similarly,

$$
\begin{aligned}
Var(Y_t) &= Var(e_t - \theta e_{t-1}) = \sigma_e^2(1+\theta^2) = \gamma_0 \\
\gamma_1 &= Cov(Y_t, Y_{t-1}) = Cov(e_t - \theta e_{t-1}, e_{t-1} - \theta e_{t-2}) \\
&= Cov(-\theta e_{t-1}, e_{t-1}) = -\theta\sigma_e^2 \\
\gamma_2 &= Cov(Y_t, Y_{t-2}) = Cov(e_t - \theta e_{t-1}, e_{t-2} - \theta e_{t-3}) \\
&= 0 \\
\rho_0 &= 1 \\
\rho_1 &= \frac{\gamma_1}{\gamma_0} = \frac{-\theta}{1+\theta^2} \\
\rho_k &= \gamma_k = 0 \qquad \text{for } k \geq 2
\end{aligned}
$$

The above are some common properties for MA(1). Similarly, I can obtain $E(Y_t), Var(Y_t), \gamma_k$ and ρ_k for MA(2) and even MA(q). For the moving average model with order 2: $Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$

$$
\begin{aligned}
E(Y_t) &= E(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}) = 0 \\
\gamma_0 &= Var(Y_t) = Var(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}) \\
&= \sigma_e^2(1 + \theta_1^2 + \theta_2^2) \\
\gamma_1 &= Cov(Y_t, Y_{t-1}) \\
&= Cov(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-1} - \theta_1 e_{t-2} - \theta_2 e_{t-3}) \\
&= Cov(-\theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-1} - \theta_1 e_{t-2}) \\
&= (-\theta_1 + \theta_1\theta_2)\sigma_e^2
\end{aligned}
$$

$$
\begin{aligned}
\gamma_2 &= Cov(Y_t, Y_{t-2}) \\
&= Cov(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-2} - \theta_1 e_{t-3} - \theta_2 e_{t-4}) \\
&= Cov(-\theta_2 e_{t-2}, e_{t-2}) \\
&= -\theta_2 \sigma_e^2 \\
\rho_0 &= 1 \\
\rho_1 &= \frac{\gamma_1}{\gamma_0} = \frac{-\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2} \\
\rho_2 &= \frac{\gamma_2}{\gamma_0} = \frac{-\theta_2}{1 + \theta_1^2 + \theta_2^2} \\
\gamma_k &= \rho_k = 0 \qquad \text{for } k \geq 2
\end{aligned}
$$

For moving average model with order q:

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

$$
\begin{aligned}
E(Y_t) &= E(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}) = 0 \\
\gamma_0 &= Var(Y_t) = \sigma_e^2(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2) \\
\gamma_k &= Cov(Y_t, Y_{t-k}) = \sigma_e^2(-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q) \qquad \text{for } k \leq q \\
\rho_k &= \frac{\gamma_k}{\gamma_0} = \frac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2} \qquad \text{for } k \leq q \\
\gamma_k &= \rho_k = 0 \qquad \text{for } k > q
\end{aligned}
$$

Autoregressive—AR

If a time series Y_t could be written as a linear combination of its own p past values, that is,

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$

where e_t is an innovation that is not explained by the past values, which means e_t is independent of $Y_{t-1}, Y_{t-2}, \ldots$. Then Y_t is called autoregressive process. Considering AR model with order 1 and order 2, $Y_t = \phi Y_{t-1} + e_t$, and $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t$, their respective $E(Y_t), Var(Y_t), \gamma_k$ and ρ_k are listed as following in Table 3.

Table 3: AR(1) and AR(2) Summaries

	AR(1)	AR(2)
$E(Y_t)$	0	0
γ_0	$\frac{\sigma_e^2}{1-\phi_2}$	$\frac{1-\phi_2}{1+\phi_2}\frac{\sigma_e^2}{(1-\phi_2)^2-\phi_1^2}$
γ_1	$\frac{\phi}{1-\phi_2}\sigma_e^2$	$\frac{\phi_1}{1+\phi_2}\frac{\sigma_e^2}{(1-\phi_2)^2-\phi_1^2}$
γ_2	0	$\frac{\phi_1^2+\phi_2-\phi_2^2}{1+\phi_2}\frac{\sigma_e^2}{(1-\phi_2)^2-\phi_1^2}$
ρ_0	1	1
ρ_1	ϕ	$\frac{\phi_1}{1-\phi_2}$
ρ_2	0	$\frac{\phi_1^2+\phi_2-\phi_2^2}{1-\phi_2}$
$\gamma_k = \rho_k (k > 2)$	0	Decay

For AR(p) model, $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$, I only give general Yule-Walker equations here instead of more details for γ_k and ρ_k, and these equations are useful in next section, parameter estimation.

For any $k \geq 1$, the Yule-Walker equation is given by:

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_k \rho_{k-p}$$

Autoregressive Moving Average—ARMA

For some time series, it could be a combination of autoregressive and moving average. Like,

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

which is called a mixed autoregressive moving average model with orders p and q respectively. I abbreviate such a time series as ARMA(p,q).

Considering ARMA(1,1) as one example,

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1}$$

obviously, Y_t has mean zero. To obtain γ_0, take variance for the previous equation and get the following result:

$$\gamma_0 = \phi^2 \gamma_0 + \sigma_e^2 + \theta^2 \sigma_e^2 + 2\phi E(Y_{t-1} e_t) - 2\phi\theta E(Y_{t-1} e_{t-1})$$

Clearly, $E(Y_{t-1}e_t)$ and $E(Y_{t-1}e_{t-1})$ need to be figured out first to solve γ_0. $E(Y_{t-1}e_t) = 0$ since e_t is independent of Y_{t-1}.

$$\begin{aligned} E(Y_{t-1}e_{t-1}) &= E[e_{t-1}(\phi Y_{t-2} + e_{t-1} - \theta e_{t-2})] \\ &= \sigma_e^2 \end{aligned}$$

Then, we have

$$\gamma_0 - \phi^2\gamma_0 = \sigma_e^2 + \theta^2\sigma_e^2 - 2\phi\theta\sigma_e^2$$
$$(1-\phi^2)\gamma_0 = \sigma_e^2(1+\theta^2-2\theta\phi)$$
$$\gamma_0 = \frac{\sigma_e^2(1+\theta^2-2\theta\phi)}{1-\phi^2}$$

To solve γ_1, we need to multiply $Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1}$ by Y_{t-1} and do similar steps, giving

$$\gamma_1 = \phi\gamma_0 - \theta\sigma_e^2$$

Combining with equation $\gamma_0 = \frac{\sigma_e^2(1+\theta^2-2\theta\phi)}{1-\phi^2}$ yields

$$\gamma_1 = \frac{(\phi-\theta)(1-\theta\phi)}{1-\phi^2}\sigma_e^2$$

Based on γ_0 and γ_1, we can obtain $\rho_0 = 1$ and $\rho_1 = \frac{(\phi-\theta)(1-\theta\phi)}{(1+\theta^2-2\theta\phi)}$. For $k \geq 2$, we need to multiply the equation $Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1}$ by Y_{t-k}, it yields

$$\gamma_k = \phi\gamma_{k-1}$$

Further, we can have

$$\gamma_k = \frac{(\phi-\theta)(1-\theta\phi)\phi^{k-1}}{1-\phi^2}\sigma_e^2$$
$$\rho_k = \frac{(\phi-\theta)(1-\theta\phi)}{(1+\theta^2-2\theta\phi)}\phi^{k-1}$$

Integrated Autoregressive Moving Average—ARIMA

A time series Y_t is said to be integrated autoregressive moving average (ARIMA(p,d,q)) if it is stationary after the dth difference with $W_t = \bigtriangledown^d Y_t$, where W_t follows ARMA(p,q). Usually, we take $d = 1$ and at most 2.

Considering ARIMA(p,1,q),

$$W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_p W_{t-p}$$

$$+e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

where $W_t = \bigtriangledown Y_t = Y_t - Y_{t-1}$. It could be written as

$$Y_t = (1 + \phi_1)Y_{t-1} + (\phi_2 - \phi_2)Y_{t-2} + (\phi_3 - \phi 2)Y_{t-3} + \ldots$$

$$+e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

Besides, $Y_t = Y_{t-1} + e_t - \theta e_{t-1}$ for IMA(1,1), $Y_t = (1 + \phi)Y_{t-1} - \phi Y_{t-2} + e_t$ for ARI(1,1), and $Y_t = (1 + \phi)Y_{t-1} + e_t - \theta e_{t-1}$ for ARIMA(1,1,1).

According to Table 1: Behavior Summary of ACF and PACF for ARMA models, I assume the time series Z follows MA(1) model from Figure 2 and the codes for ACF is in [11], which also means time series Y follows IMA(1,1), but I still need to obtain estimates of parameters.

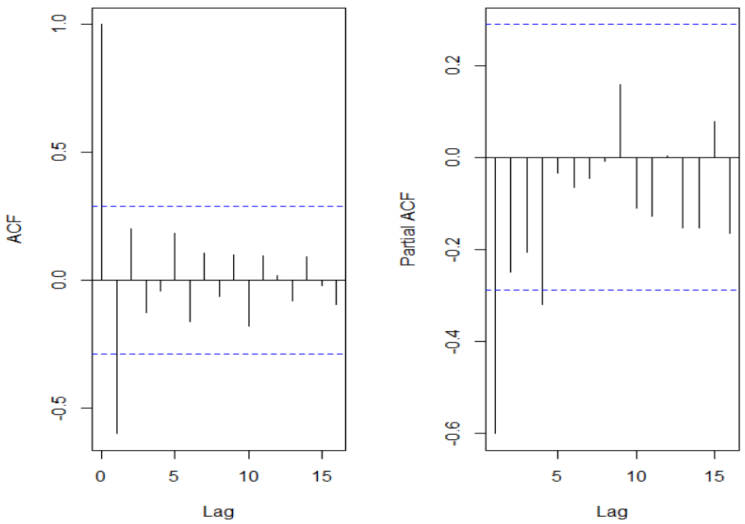

Figure 2: ACF and PACF for Jan 1st (Z value)

1.5 Parameter Estimation

The Method of Moments

The method of moments is one of the easiest method for obtaining parameter estimates, although it's not the most sufficient.The essence of method of moments is to equate sample moments to corresponding theoretical moments, and then solving the equations to obtain estimates of unknown parameters.

For AR(1) model, based on Table 2, $\rho_1 = \phi$, if I equate ρ_1 to be r_1, the lag 1 sample autocorrelation. Then, ϕ is estimated by

$$\hat{\phi} = r_1$$

Now, considering AR(2) model, $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t$, by the Yule-Walker equations

defined previously, taking $k = 1, 2$ gives

$$\rho_1 = \phi_1 + \rho_1\phi_2$$

$$\rho_2 = \rho_1\phi_1 + \phi_2$$

Replacing ρ_k by r_k gives

$$r_1 = \phi_1 + r_1\phi_2$$

$$r_2 = r_1\phi_1 + \phi_2$$

Then, solving the above equations to obtain estimations for ϕ_1 and ϕ_2 given by

$$\hat{\phi_1} = \frac{r_1(1-r_2)}{1-r_1^2} \text{ and } \hat{\phi_2} = \frac{r_2-r_1^2}{1-r_1^2}$$

Running code $ar(data, order.max, AIC, method)$ in R language called gives Method-of-Moments estimates.

Considering MA(1) model, $Y_t = e_t - \theta e_{t-1}$, and we also know that $\rho_1 = -\frac{\theta}{1+\theta^2}$. Equating $\rho_1 = r_1$ yields a quadratic equation $r_1\theta^2 + \theta + r_1 = 0$.
If $|r_1| < 0.5$, there will be real roots, and

$$\hat{\theta} = \frac{-1 \pm \sqrt{1-4r_1^2}}{2r_1}$$

If $|r_1| = 0.5$, we will have $\hat{\theta} = \pm 1$. If $|r_1| > 0.5$, there is no real roots, and it also means I cannot get estimates for MA(1) model, so the MA(1) model is probably not good.
For higher order MA models, $Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_1 e_{t-q}$, similarly I have

$$\gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma_e^2$$

and

$$\rho_q = \begin{cases} \frac{-\theta_k+\theta_1\theta_{k+1}+\theta_2\theta_{k+2}+\cdots+\theta_q\theta_{k+q}}{1+\theta_1^2+\theta_2^2+\cdots+\theta_q^2}, & \text{for } k = 1,2,\ldots,q \\ 0, & \text{for} k > q \end{cases}$$

I can use the above equations to obtain estimates for θ.

Considering ARMA(1,1), $Y_t = \phi Y_t + e_t - \theta e_{t-1}$. We know

$$\rho_k = \frac{(1-\theta\phi)(\phi-\theta)}{1-2\theta\phi+\theta^2}\phi^{k-1} \quad \text{for} \quad k \geq 1$$

The estimation of $\hat{\phi}$ is $\frac{r_2}{r_1}$, because $\frac{\rho_2}{\rho_1} = \phi$. Besides, when $k = 1$,

$$r_1 = \frac{(1-\theta\phi)(\phi-\theta)}{1-2\theta\phi+\theta^2}\phi$$

it's a quadratic equation with θ. After calculation, if $\Delta \geq 0$, the estimate of $\hat{\theta}$ is given by :

$$\hat{\theta} = \frac{(\hat{\phi}^2+1-2\hat{\phi}r_1) \pm \sqrt{\Delta}}{2(\hat{\phi}-r_1)} \quad \text{where} \quad \Delta = (2\hat{\phi}r_1 - 1 - a^2)^2 - 4(\hat{\phi}-r_1)^2$$

Maximum Likelihood

For any time series, $Y_1, Y_2, \ldots, Y_n$, the likelihood function L is defined by the joint pdf with unknown parameters, $L = f(\underline{y}|\hat{\phi}'s, \hat{\theta}'s, \hat{\mu}, \hat{\sigma_e^2})$. The maximum likelihood estimators, $\hat{\phi}, \hat{\theta}, \hat{\mu}, \hat{\sigma_e^2}$, are obtained by maximizing the likelihood function L.

Least Squares

The conditional log-likelihood function l_* in [7, 8] is defined as

$$l_* = -\frac{n}{2}ln(\sigma_e^2) - \frac{S_*(\hat{\phi}'s, \hat{\theta}'s, \hat{\mu})}{\sigma_e^2}$$

where $S_*(\hat{\phi}'s, \hat{\theta}'s, \hat{\mu}) = \sum_{t=1}^{n} e_t^2(\hat{\phi}'s, \hat{\theta}'s, \hat{\mu})$ is called conditional sum-of-squares function in [7] and [8].

Running code $arima(data, order, method)$ in R, with methods CSS and ML, gives corresponding least squares estimates and maximum likelihood estimates.

Table 4 shows different estimates with the above three different methods for Jan 1st, Jan 5th and Jan 17th. It seems that there are not much difference among different estimates methods, but Method-of-Moments doesn't do very well for the parameter estimation. In this project, I use Maximum Likelihood Estimates or Least Squares Estimates for our models.

Table 4: Parameter Estimation for Different Models

Parameter	Method-of-Moments Estimates	Least Squares Estimates	Maximum Likelihood Estimates
	Jan 1st-IMA(1,1) Model		
θ	NA	1.1076	1
	Jan 5th-ARI(4,1) Model		
ϕ_1	-0.8907	-0.9966	-0.9377
ϕ_2	-0.5658	-0.6117	-0.6219
ϕ_3	-0.5202	-0.5293	-0.5848
ϕ_4	-0.3465	-0.3573	-0.3783
	Jan 17th-ARMA(1,1,1)		
ϕ	-0.413	-0.3516	-0.3183
θ	0.428	0.7473	0.7532

1.6 Model Diagnostics

Model diagnostics is proposed to test the goodness of fit of a model. If the fit is poor, it means appropriate modifications or improvements are necessary to take. For this project, a main approach of diagnostics, analysis of residuals from the fitted model, is proposed. There is one important criteria of model specification needed to be proposed first. Akaike's Information Criteria (AIC) [6, 11] is the most studied approach to model specification.

$$AIC = -2\log(\text{maximum likelihood}) + 2(p+q+1)$$

where p, q are orders from ARMA model. Higher orders p, q will offer penalty for AIC, but $-2log(\text{maximum likelihood})$ will make up for it greatly. In R, $auto.arima(data)$ [6] automatically gives a model with least AIC whose residuals will be analysed for model diagnostics.

There are three main approaches for residuals analysis: plots of residuals, the normality of residuals, and sample ACF for residuals. Checking residuals is usually the first step. If most residuals are approaching to zero, the model is not bad. Besides, it's also necessary to check the normality of the residuals. There are various methods: qqnorm, qqline, histogram and Shapiro-Wilk normality test (Shapiro.test() in R). Finally, for a good model, the sample autocorrelations are approximately uncorrelated.

With MLE method, lake level data of Jan 1st from 1971 to 2017 follows IMA(1,1) model with least AIC= 313.22, Table 5 gives details of the model. The standardized residuals plot supports the model in Figure 3, since no trends are present and most residuals are around zero. Besides, the quantile-quantile plot shows the points follow the straight line. In addition, the Shapiro-Wilk normality test yielding a result, I get $W = 0.98031$ and corresponding p-value is $0.6186 > 0.05$. Thus, we cannot reject the normality according to the test.

Table 5: Maximum Likelihood Estimation of IMA(1,1) Model for Jan 1st

Coefficients:	IMA(1,1)	Intercept
	1.000	0.0779
s.e.	0.089	0.0704

σ^2 estimated as 42.82 :log likelihood =-153.61,aic = 313.22

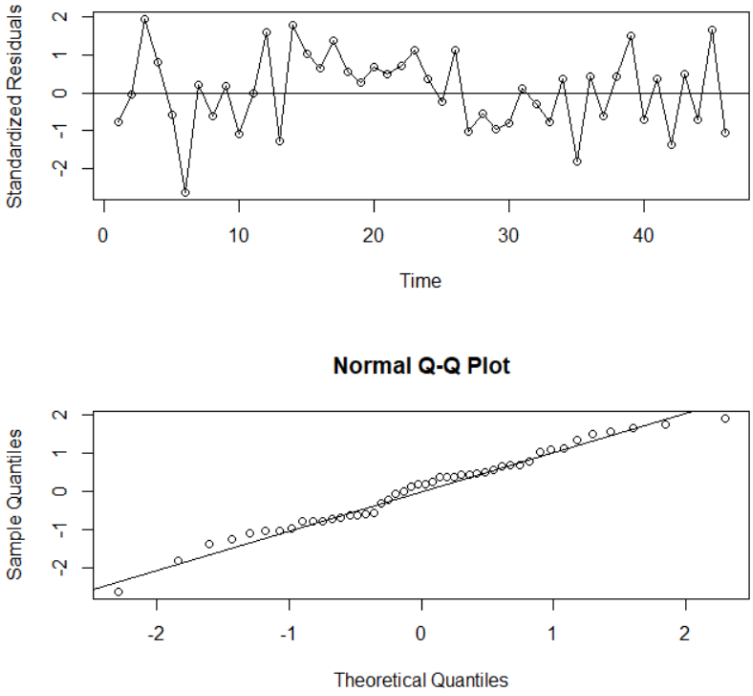

Figure 3: Residuals Analysis for IMA(1,1) for Jan 1st

Finally, from Figure 4, the ACF residuals shows no correlation since they are all inside the bounds. In other words, we can forecast the lake level on Jan 1st, 2018 based on IMA(1,1) in next section.

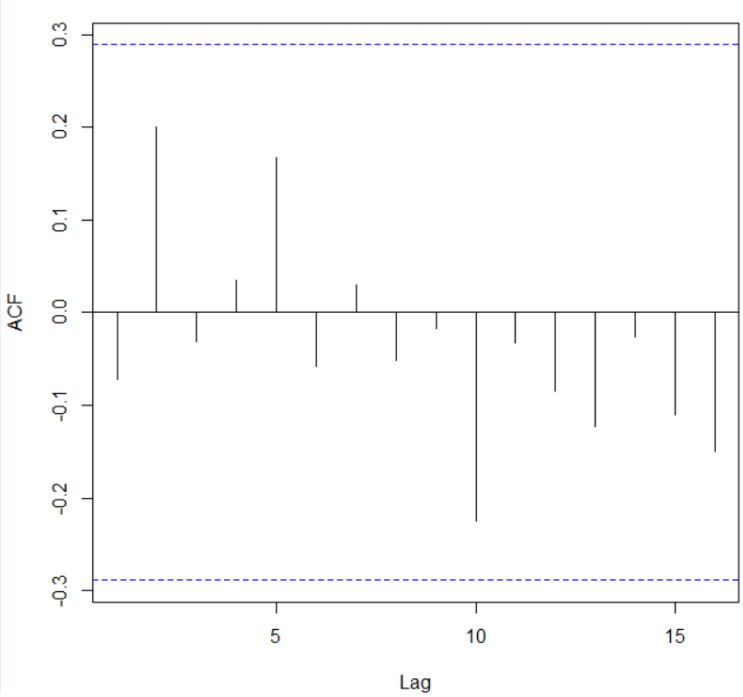

Figure 4: ACF Residual for IMA(1,1) for Jan 1st)

1.7 Forecasting

The main purpose of building a model for a time series is to make precise predictions for the series in the future. In my thesis, there are three modeling methods: regression method, functional ARIMA method and multiplicative seasonal ARIMA method. For all three methods, minimizing square error is the main principle for prediction.

Next, I will use IMA(1,1) model to make predictions as one example. Figure 5 displays the forecasted value of the time series together with upper and lower 95% bounds. With Table 5, we know $Y_t = 0.0779 + e_t - e_{t-1} + Y_{t-1}$. We can have one-step-ahead forecast for the IMA(1,1) model expressed as

$$\hat{Y}_t(1) = \mu - \theta e_t = 0.0779 + e_t + Y_{t-1}.$$

We can have one-step-ahead forecast from R language as 1113.005458. With similar steps,

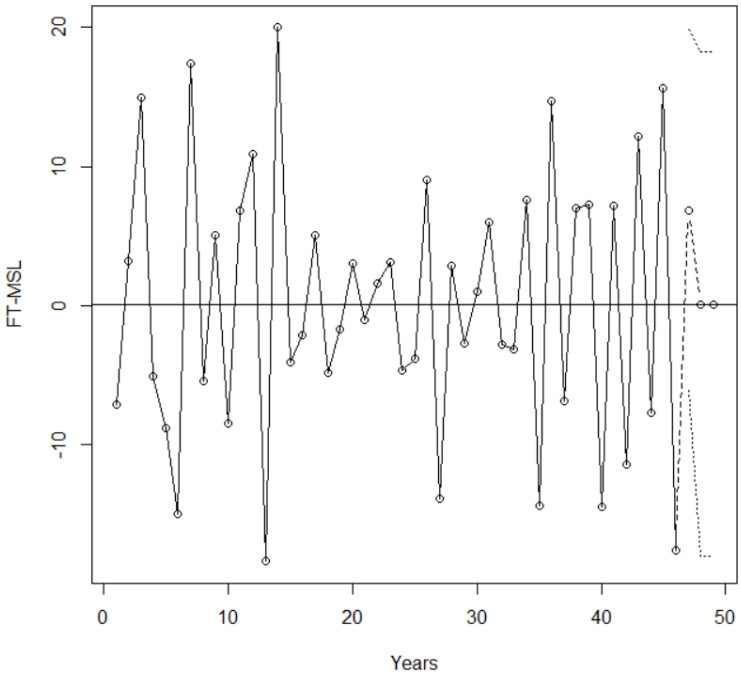

Figure 5: Forecast for IMA(1,1) for Jan 1st, 2018

we can get models and prediction for everyday in 2018, which will be introduced in detail in section application to lake level of beaver lake with functional ARIMA method.

2. APPLICATION TO WEATHER DATA IN NYCP

I collected daily maximum temperature data for New York Central Park from 1870-2017 [1], and try to do analysis and make predictions based on what I mentioned in the introduction. Obviously, there are a bunch of various temperature for the park, and I consider the maximum temperature per day. Noticing there are some missing values during Sep 10th - Dec 31th in 2011, and Jun 28th - Dec 31th in 1949. In order to make our data complete and conduct our analysis better, it's absolutely necessary for us to fill these missing values. To improve accuracy of data I will fill in, I will make prediction for per day's maximum temperature based on all maximum temperature values of the same day prior to the present years. That is to say, if I want to replace NAN on Jun 28th in 1949 with a number, called A, then I need to make a prediction for A based on a time series which is made up with all values on Jun 28th from 1870 to 1948. For each time series model, there is a ARIMA model with least AIC, and I will give a prediction for number A according to the model, which is how I make the data complete. Next, I will use three different methods: regression method, functional ARIMA method, multiplication seasonal ARIMA method, to do analysis for weather data in New York Central Park.

2.1 Regression Method

For regression method, I consider the data as three parts:

$$Y_t = M_t + S_t + X_t + \epsilon$$

where M_t is a polynomial of the deterministic or trend, S_t represents seasonality, X_t is a stationary ARIMA model, and ϵ is the error term.

In order to have a good overview for the maximum temperature trend in New York Central Park from 1870 to 2017, I plot its temperature in the first and last 10 years, and

maximum temperature from 2013 to 2017 as examples in Figure 6 and Figure 7 respectively. Note, the unit of maximum temperature is 0.1°C.

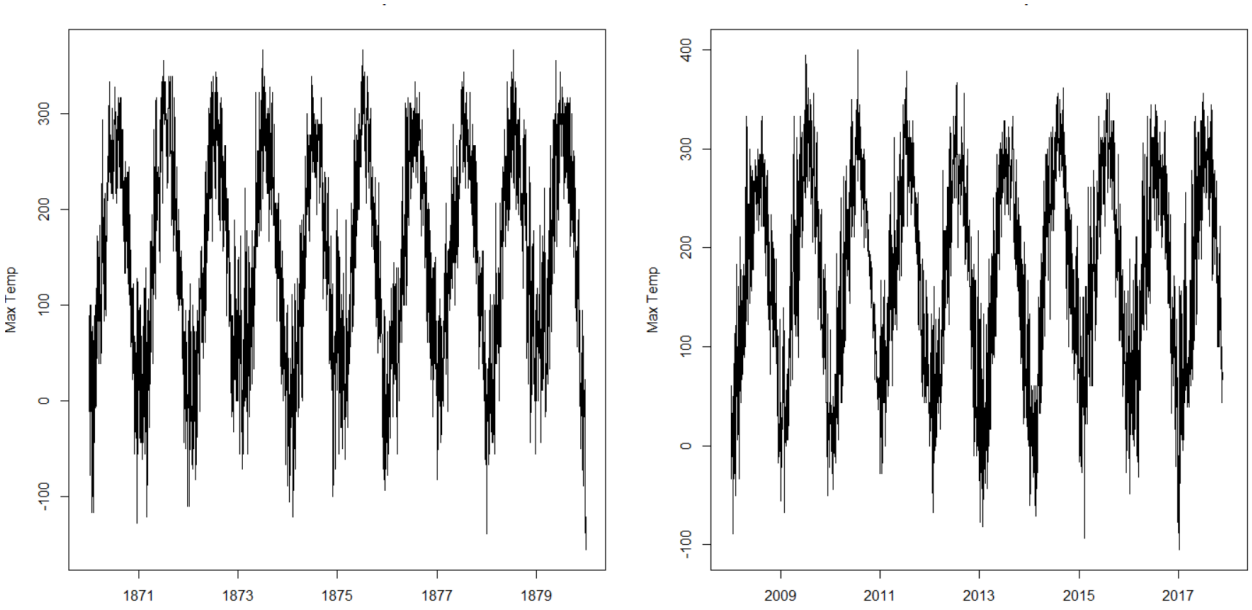

Figure 6: Graph for NYCP 1870-1879 and 2008-2017

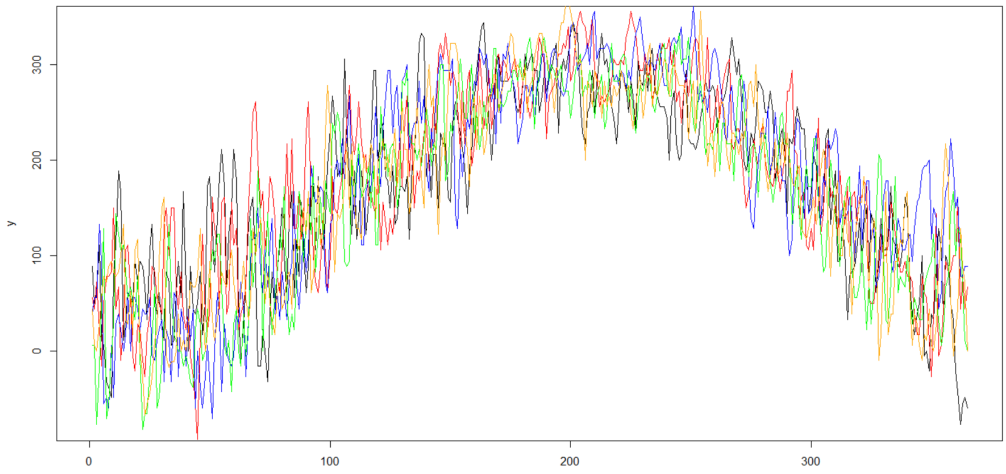

Figure 7: Graph for NYCP 2013-2017

From the above graphs, the max temperature trend is approximately harmonic with slight differences for each year. Thus, I will deal with it as a harmonic model via a trend with regression method, where

$$M_t = 145.7 + 1.104 \times 10^{-3} \times t - 9.52 \times 10^{-9} \times t^2$$

S_t is a harmonic series,

$$S_t = -0.0737 - 122.7512cos(2\pi t) - 3.5284cos(4\pi t) - 45.0796sin(2\pi t) - 1.0359sin(4\pi t)$$

With models of M_t and S_t, I tried plotting a fitting graph with observations from 2007 to 2017 as displayed in Figure 8.

However, the model doesn't fit well for the extrema. Thus, I need to model the residuals

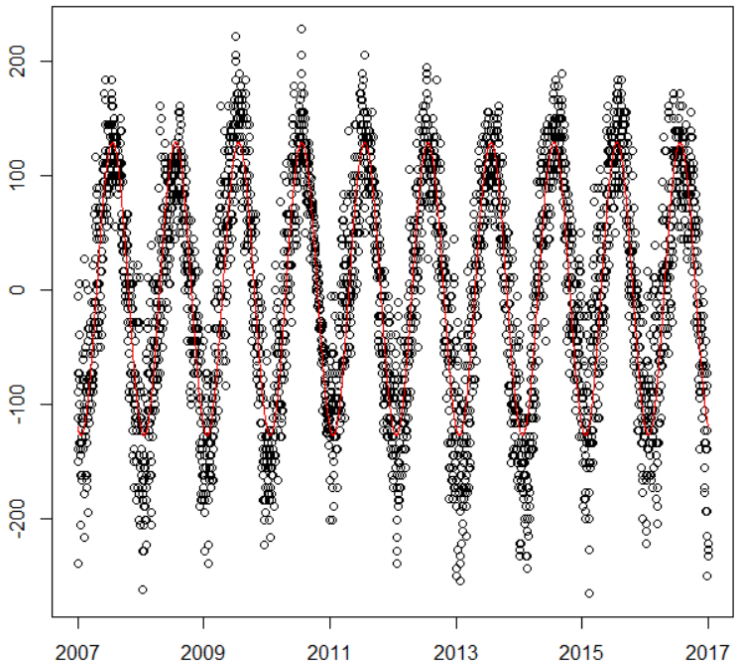

Figure 8: Improved Cosine Trend for NYCP 2007-2017

to minimize error term. Mean of $X_t = -9.592372 \times 10^{-17}$, it's almost zero. After modeling,

X_t follows a stationary ARMA(2,3) with parameters as followed in Table 6.

Table 6: ARMA(2,3) for X_t

	ar1	ar2	ma1	ma2	ma3
estimate	1.4023	-0.4303	-0.7788	-0.1750	0.0439
s.e.	0.0395	0.0340	0.0400	0.0117	0.0148

I forecast daily values for 2018 with $Y_t = M_t + S_t + X_t + \epsilon$, and there is a comparison between real values and the forecasted in Figure 9.

From the comparison, our predictions work well for most dates except some days in February.

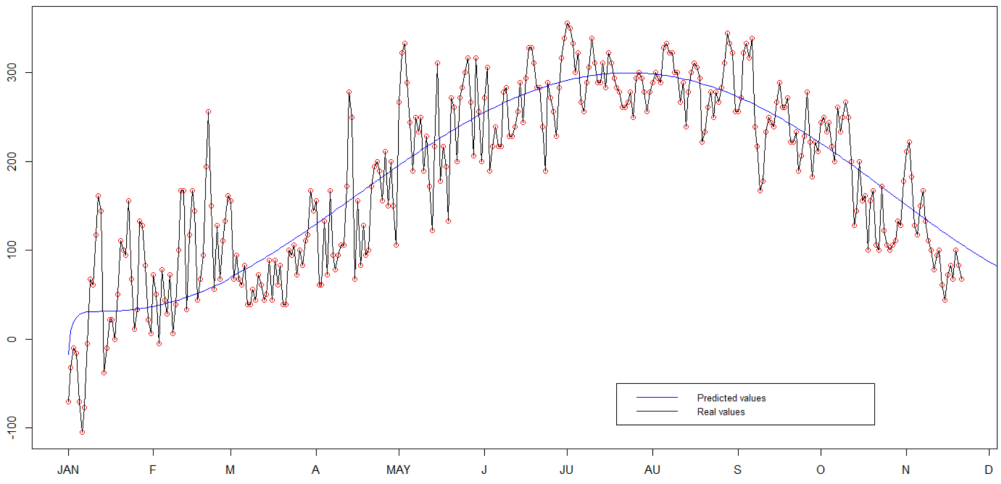

Figure 9: Prediction VS Observations NYCP

To explore what happened in February, I checked weather data for the past 20 years, and found that the maximum temperature values in some days of February reach the highest in 2018, which is why I cannot forecast perfectly for this period based on the past values. In next chapter, for lake level data, something similar happened as well. A detailed forecasting of maximum temperature in NYCP from Jan 1st to Feb 11th in 2018 with this method is

given in Appendix A.

2.2 Functional ARIMA Method

For functional ARIMA, assuming

$$Y_i(t) = V_i(t) + \epsilon_i$$

where $t = 1, 2, \ldots, 365$, and $i = 1870, 1871, \ldots, 2017$.

After running *adf.test()* in R, for any specific year i, $Y_i(t)$ is stationary after lag 1. Besides, data on most dates in the past 148 years, follow similar ARIMA models displayed in Table 7. EVen though there are 15 different models for each year, there are not much differences between their orders, and the model with more parameters will improve its accuracy.

Similarly, with any specific t, $Y_i(t)$ is stationary after lag 1, and there are 20 distinct ARIMA models in total in Table 8. For example, there are 165 dates following IMA(1,1) model. Thus, it's hopeful for me to seek a general model suitable for any date.

Table 7: ARIMA Models for Each Year 1870-2017

Models	IMA(1,2)	IMA(1,3)	ARIMA(1,1,1)	RIMA(1,1,2)
Number	14	22	13	22
Models	RIMA(1,1,3)	ARIMA(2,1,1)	ARIMA(2,1,2)	ARIMA(2,1,4)
Number	8	30	14	2
Models	ARIMA(3,1,1)	ARIMA(3,1,2)	ARIMA(3,1,3)	ARIMA(3,1,4)
Number	4	5	1	2
Models	ARIMA(4,1,1)	RIMA(4,1,4)	ARIMA(4,1,5)	
Number	6	3	1	

Table 8: NYCP Lake ARIMA Models Summary

Model	Number	Dates
IMA(1,1)	165	The rest
IMA(1,2)	18	03/22,04/10,04/21,05/06,05/15,06/08,07/02,07/09 07/21,07/29,07/30,08/08,10/02,10/08,10/09,11/01 11/14,12/11
IMA(1,4)	1	10/05
ARIMA(1,1,1)	14	01/03,05/17,07/05,07/10,07/20,07/25,09/04,09/05 09/12,09/21,10/06,10/13,11/06
ARIMA(1,1,2)	9	06/11,07/19,07/24,09/13,11/04,11/19,11/31,12/01 13/06
ARIMA(2,1,0)	1	09/17
ARIMA(2,1,1)	27	01/11,01/29,02/06,02/11,02/19,02/24,02/25,04/14 04/15,04/20,05/03,05/14,05/24,05/25,05/30,08/1 08/20,08/24,08/25,09/08,09/20,09/27,10/24,10/25 11/27,12/21,12/29
ARIMA(2,1,2)	6	02/12,02/28,03/05,05/16,06/10,12/07
ARIMA(2,1,3)	2	07/16,11/16
ARI(3,1)	14	01/20,02/26,05/21,06/13,06/20,06/21,06/28,07/14 09/18,09/25,09/29,10/20,10/27,11/12
ARIMA(3,1,1)	23	01/04,01/05,01/26,02/01,02/05,02/16-02/18,02/23 03/12-03/14,04/07,06/27,07/22,08/03,08/10,09/03 09/28,11/05,11/08,12/26,12/30
ARIMA(3,1,2)	6	03/16,06/18,07/18,07/28,12/05,12/10
ARIMA(3,1,3)	1	11/20
ARIMA(4,1,0)	15	01/06-01/08,01/19,01/28,04/30,05/26,06/14,06/24 08/12,09/22,10/10,10/26,12/12,12/14
ARIMA(4,1,1)	18	01/30,03/11,03/29,04/17,04/22,05/18,05/28,06/04 06/05,08/18,09/06,09/14,11/02,11/03,11/13,11/25 12/20,12/24
ARIMA(4,1,2)	1	08/05
ARI(5,1)	29	01/02,01/16-01/18,01/23,01/27,01/31,02/13,03/03 05/08,05/09,05/19,06/02,06/17,06/25,06/29,07/03 07/04,07/12,07/13,08/07,08/11,08/22,10/01,10/03 10/04,11/23,12/13,12/19
ARIMA(5,1,1)	12	01/09,02/14,03/15,03/18,05/10,05/12,05/20,06/12 08/06,08/26,11/24,12/14
ARIMA(5,1,2)	2	01/14,06/19
ARIMA(5,1,3)	1	10/21

According to Table 7 and Table 8, in the general model ARIMA(p,d,q), undoubtedly, $d = 1$, since the time series with any specific t or i is stationary with after taking the first difference. The highest order for p is 5, and it accounts for a lot. Thus, I could take $p = 5$. For q value, the highest order is also 5. However, it works in only one model, and it seems that $q = 2$ and $q = 3$ accounts for a large proportion. Clearly, $q = 3$ is a better choice than $q = 2$ since it will improve accuracy with one more parameter. Thus, I consider a general model ARIMA(5,1,3) or ARIMA(5,1,5). In order to choose a more suitable one, I compared them with maximum temperature in most dates and most years randomly. Here, a comparison result for Mar 2nd from 1870 to 2017 is given in Table 9.

Table 9: ARIMA(5,1,5) and ARIMA(5,1,3) for Mar 2nd

	ar1	ar2	ar3	ar4	ar5	ma1	ma2	ma3
estimate	-0.6756	-0.6174	-0.0446	-0.1009	-0.1382	-0.2727	-0.0665	-0.5568
s.e.	0.3959	0.3912	0.1215	0.1122	0.0962	0.3972	0.5999	0.3697
σ^2 estimated as 2981,log likelihood=-796.43,aic=1608.85								
	ar1	ar2	ar3	ar4	ar5	ma1	ma2	ma3
estimate	-0.4081	-0.4346	-0.3821	-0.9442	-0.1274	-0.5295	0.0069	-0.0152
s.e.	0.0820	0.0507	0.0514	0.0541	0.0886	0.0501	0.0528	0.0379
	ma4	ma5						
estimate	0.6049	-0.9212						
s.e.	0.0.0528	0.0533						
σ^2 estimated as 2654,log likelihood=-791.67,aic=1603.34								

The automatically given model by R is IMA(1,1) with $AIC = 1600.62$. For models, ARIMA(5,1,3) and ARIMA(5,1,5), though there are some differences among parameters, the discrimination among AIC values is small and ARIMA(5,1,5) has the least AIC, which works for most dates and years. Thus, model ARIMA(5,1,5) becomes the general one. Based on the general model, I forecast for 2018 with 95% confidence level and make a comparison with observations in Figure 10.

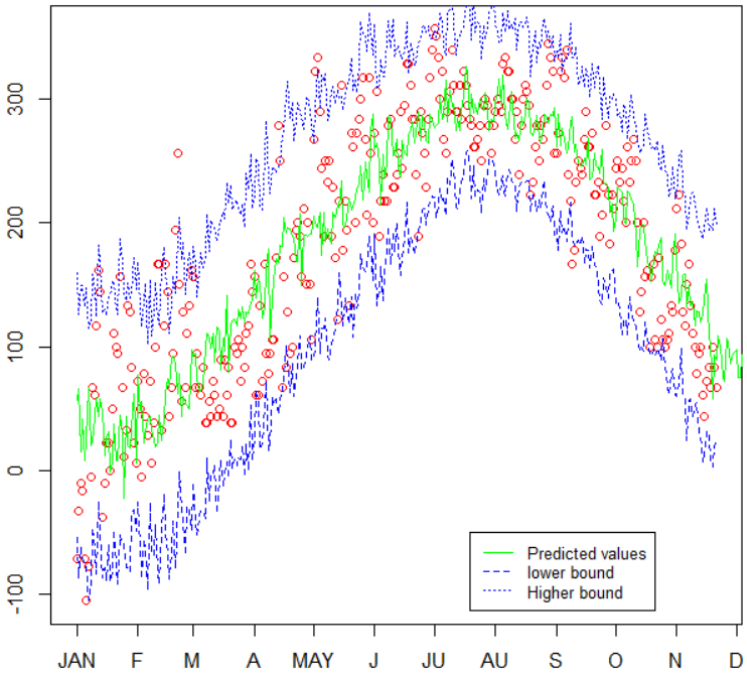

Figure 10: Prediction VS Observations NYCP Func-ARIMA

From the comparison, almost all observations are inside the bounds except 5 points. Thus, we can believe the model and forecasting.

2.3 Multiplicative Seasonal ARIMA Method

Multiplicative ARIMA model is to combine seasonal and non-seasonal ARIMA models, which not only considers seasonal effects on the current but also neighborhood effects on

the current data. A multiplicative ARMA is denoted as ARIMA $(p,q)\times(P,Q)_s$ model with seasonal period s. In R, s is required to be less than 350. Actually, when $s > 200$, the prediction is less accurate. Thus, the data of daily maximum temperature in New York Central Park is divided into weekly data where the average is taken for per week, and the new data is a seasonal time series with period 52. The Table 10 shows ARIMA$(4,0,1)\times(4,0,1)_{52}$ for the new time series.

Table 10: Multiplicative Seasonal Model NYCP

	ar1	ar2	ar3	ar4	ma1	sar1	sar2	sar3
estimate	0.5481	-0.0224	0.0165	0.0310	-0.3140	0.9375	-0.0027	0.0172
s.e.	0.2179	0.0526	0.0176	0.0153	0.2192	0.0119	0.0156	0.0156
	sar4	sma1	intercept					
estimate	0.0467	-0.9083	164.7003					
s.e.	0.0118	0.0058	NaN					

From the EACF of residuals in Table 11, the residuals follows ARMA(0,0), which means there is no trend for the residuals. Besides, The Ljung-Box test [2] for this model gives a chi-square of 0.014739 with 1 degree of freedom, and p-value of 0.9034. The null hypothesis of Ljung-Box is that the model does not show lack of fit; the alternative hypothesis is that the model is lack of fit. The test result indicates the model is fine and has captured the dependence in the time series. I also checked histogram and qqnorm for residuals, there is nothing wrong. Thus, I can make predictions for 2018 based on the model.

Figure 11 displays forecasting of weekly average maximum temperature in 2018 with 95% forecast limits. The forecast limits are quite close to the fitted values due to small error variance. In addition, the forecasting has captured the trend and works well for most dates. From the comparison with observations, there is not much difference, and we can believe the

model and forecasting. A detailed forecasting of maximum temperature in 2018 with this method is given in Appendix B.

Table 11: Extended ACF (EACF) of Residuals from Multiplicative ARIMA

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	x	0	0	0	0	0	0	0	0	0	0	0	0	0
2	x	x	0	0	0	0	0	0	0	0	0	0	0	0
3	x	x	x	0	0	0	0	0	0	0	0	0	0	0
4	x	x	x	0	0	0	0	0	0	0	0	0	0	0
5	x	x	x	x	x	0	0	0	0	0	0	0	0	0
6	x	x	x	x	x	x	0	0	0	0	0	0	0	0
7	x	x	0	x	x	0	0	0	0	0	0	0	0	0

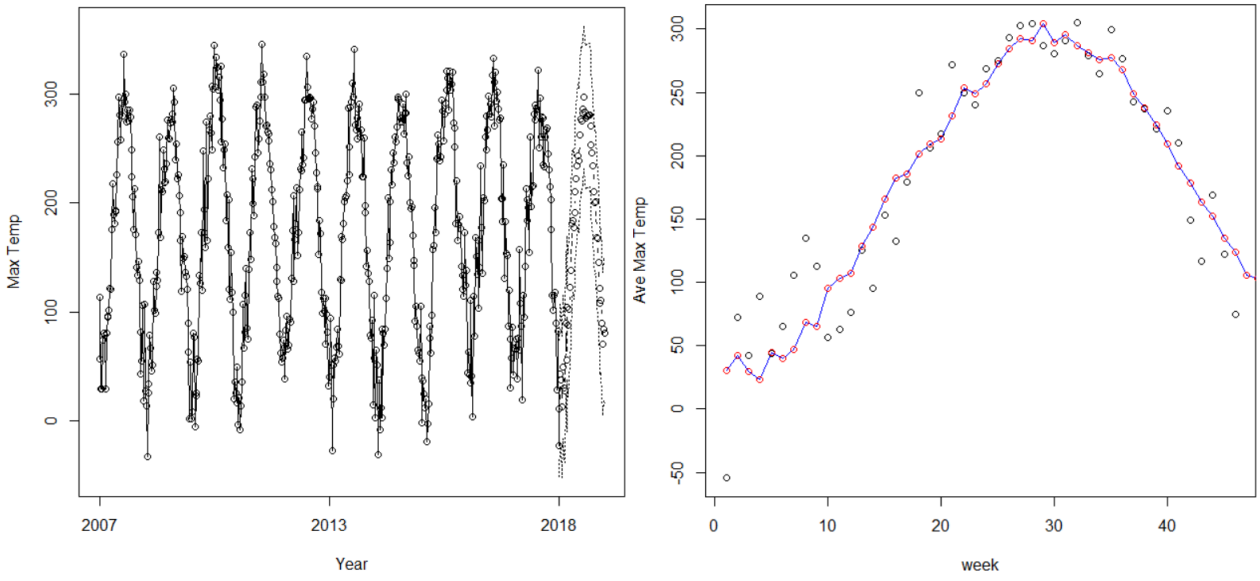

Figure 11: Forecasting and comparison for NYCP Multi Model

3. APPLICATION TO LAKE LEVEL OF BEAVER LAKE

Hourly lake level data of Beaver Lake was collected from 1987 to 2017, but there is little difference among data within one day. Besides, I only obtain data of 7:00 am everyday starting from 1971. Thus, to have data over more years for a time series model, observations of 7:00 am each day from 1971 to 2017 are chosen for modeling and forecasting. Note, the unit of lake level is denoted as *FT-MSL*: feet-mean sea level.

From Figure 12, we can see that the $FT - MSL$ values for each year approximately follow a similar path with different amplitude. Like 2016, the $FT - MSL$ in January reaches the maximum than any other years; in 1977, the $FT - MSL$ in January reaches the minimum than any other years.

If I connect the data of 47 years together, I will get the following graph of a time series data

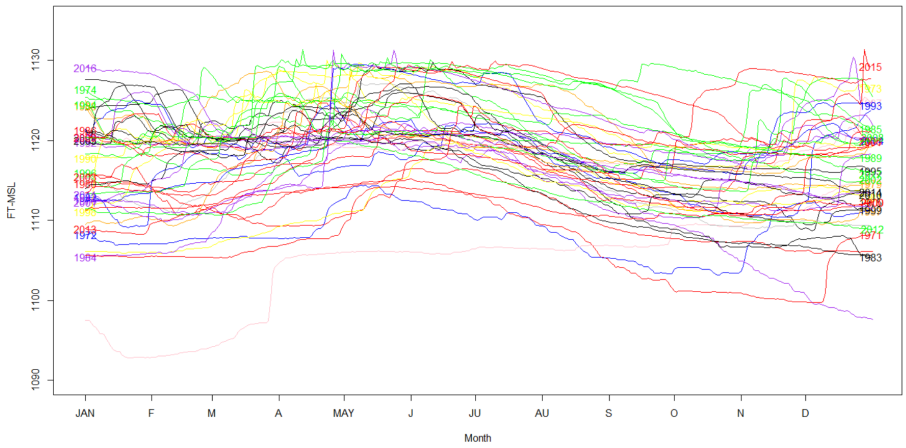

Figure 12: Beaver lake yearly (1971-2017)

as displayed in Figure 13. Just as I did for weather data previously, three different methods will be applied for modeling and forecasting for 2018.

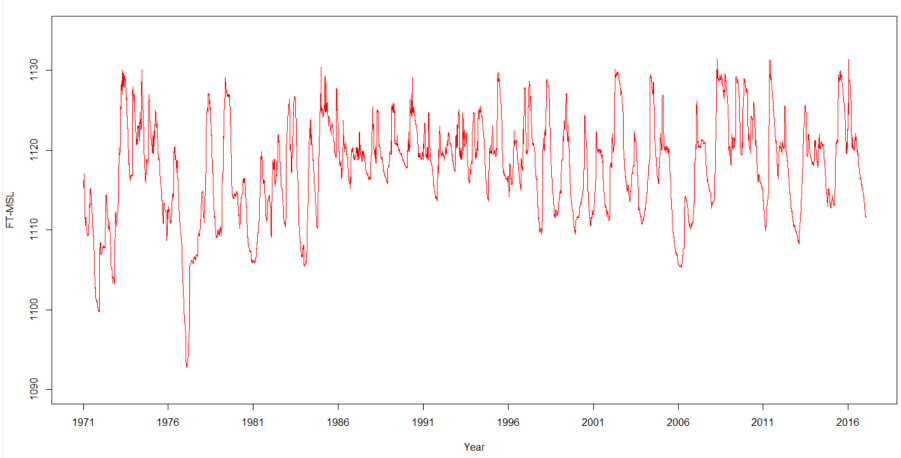

Figure 13: Beaver lake (1971-2017)

3.1 Regression Method

Similarly, assuming

$$Y_t = M_t + S_t + X_t + \epsilon$$

where $t = 1, 2, \ldots, 365$ for the regression method, M_t is still a polynomial for deterministic or trend. To determine the order of M_t, I try to choose one whose most residuals are around zero, and the mean of residuals is almost zero, among all polynomials with orders from 1 to 12. From Figure 14, when the order of polynomial is 7, the mean of residuals is almost zero, and there is not much difference among the residuals plots when the order is greater than 7. Thus, I choose the polynomial with order 7 for M_t.

$$M_t = 1106 + 2.19 * 10^{-2} * t - 1.619 * 10^{-5} * t^2 + 5.114 * 10^{-9} * t^3 - 7.947 * 10^{-13} * t^4$$

$$+6.433 * 10^{-17} * t^5 - 2.607 * 10^{-21} * t^6 + 4.183 * 10^{-26} * t^7$$

Though the Dickey-Fuller Test says the residuals are stationary, from the plot, we know it still contains seasonality. Thus, I need to find a seasonal model. Assuming to set $Z_t = S_t + X_t$, here, I use $\hat{S}_t$to estimate S_t. $\hat{S}_t$ equals the average value of every-day's residuals in the past

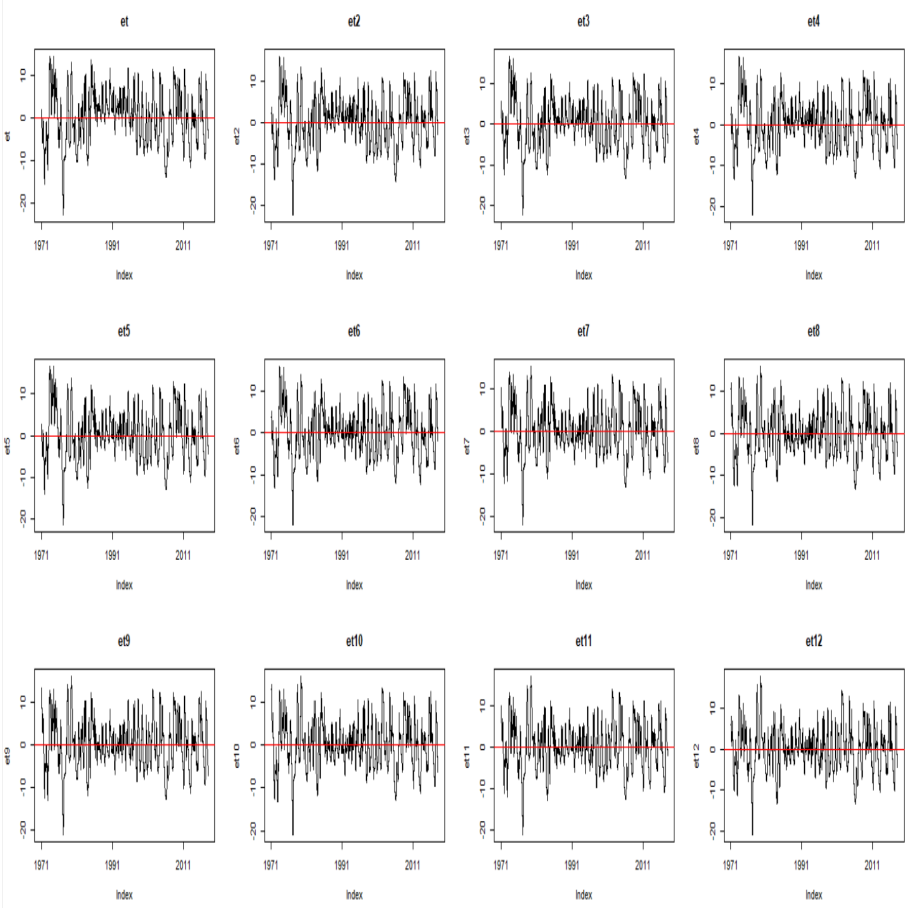

Figure 14: Residuals for Polynomials with order 1-12

47 years, which is called a seasonal means model, and $X_t = Z_t - \hat{S}_t$.

$$\hat{S}_t = \begin{cases} \beta_1, & \text{for } t = 1 \text{ from 1971 to 2017} \\ \beta_2, & \text{for } t = 2 \text{ from 1971 to 2017} \\ \vdots & \\ \beta_{365}, & \text{for } t = 365 \text{ from 1971 to 2017} \end{cases}$$

After removing $\hat{S}_t$, there will be a new plot for X_t in Figure 15. Fortunately, it follows MA(5) as a stationary time series.

With the equation, $Y_t = M_t + S_t + X_t + \epsilon$, I make predictions for daily lake level for 2018

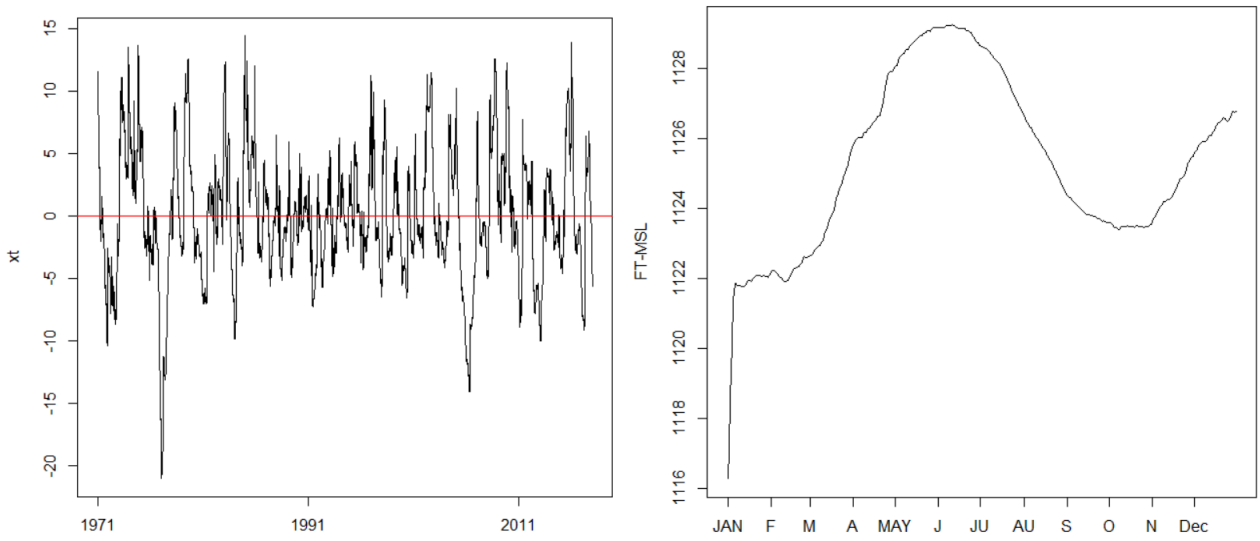

Figure 15: Plot of X_t and Forecasting of Lake Level with Regression Method

as displayed in Figure 15. This forecasting successfully captures the trend and forecasts well for most dates except some earlier time and the end of year. A detailed forecasting of Beaver Lake from Jan 1st to Feb 11th in 2018 with this method is given in Appendix A.

3.2 Functional ARIMA Method

Assuming

$$Y_i(t) = X_i(t) + \epsilon_i$$

where $i = 1971, 1972, \ldots, 2017$ and $t = 1, 2, \ldots, 365$. As I did previously for the data of NYCP, $X_i(t)$ follows ARIMA(p,1,q), where p, q vary from t. But there are also some similarities between these models as shown in Table 12. There are 150 dates which follow IMA(1,1). For this section, I make predictions for year 2018 using their own models for distinct t instead of seeking a general model, noting that there is little differences of forecasting between these two methods though.

Using corresponding ARIMA(p,1,q) for various t, $t = 1, 2, \ldots, 365$, there is a forecasted corresponding value for day t in 2018. All of forecasted values form a graph of prediction as displayed in Figure 16.

Table 12: Beaver Lake ARMA Models Summary

Model	Number	Dates
IMA(1,1)	150	01/01-01/04,02/01-03/21,04/12,06/18-07/01,08/20-10/09,12/01-12/27,12/29-12/31
IMA(1,2)	1	11/29
IMA(1,3)	2	04/19,04/21
ARI(1,1)	41	04/14-04/17,04/20,04/22-05/10,05/28,10/15-10/21,11/04-11/08,11/25-11/28,11/30
ARI(4,1)	13	01/05-01/16,12/28
ARIMA(1,1,1)	50	01/17-01/31,10/10-10/14,10/22-11/03,11/09-11/24
ARIMA(2,1,1)	23	03/22-04/11,04/13,04/18
ARIMA(2,1,2)	6	06/10,06/14,07/18,07/22,07/29,08/01
ARIMA(3,1,2)	15	06/08-06/09,06/12-06/13,06/15-06/17,07/16,07/17,07/19-07/21,07/24,08/16,08/17,
ARIMA(4,1,1)	44	05/11-05/27,05/29-06/07,06/11,07/02-07/15,07/23,07/26
ARIMA(4,1,2)	9	08/09-08/15,08/18,08/19
ARIMA(5,1,1)	11	07/25,07/27,07/28,07/30,07/31,08/02-08/05,08/07,08/08
ARIMA(5,1,2)	1	08/06

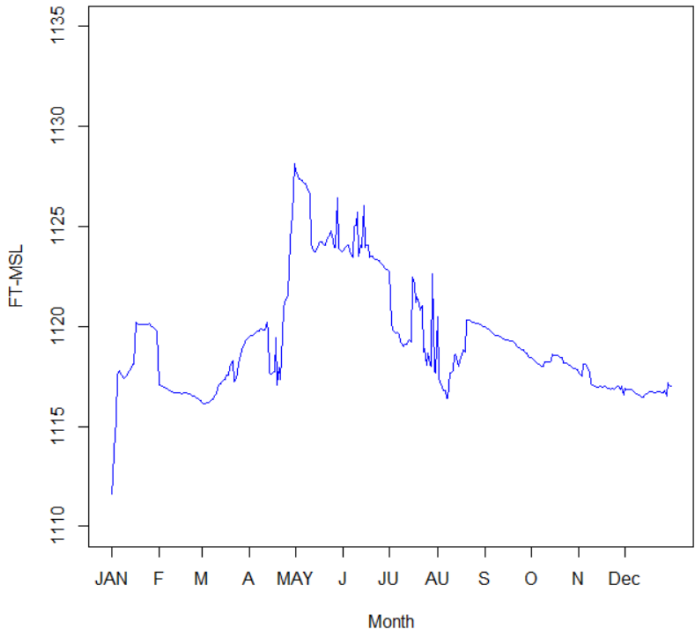

Figure 16: Forecasting of Lake Level with Functional ARIMA Method

3.3 Multiplicative ARIMA Method

As I did for weather data in New York Central Park previously, daily data of lake level in Beaver Lake is divided into weekly data where the average is taken to ensure period s is allowed in R. The new time series of weekly data follows a multiplicative $ARIMA(1,1,2) \times (1,1,2)_{52}$ with parameters as followed in Table 13:

Table 13: Parameters for Multiplicative ARIMA Model

	a1r	ma1	ma2	sar1	sma1	sma2
estimate	0.4605	-0.0194	-0.0882	-0.4313	-0.4873	-0.4063
s.e.	0.1843	0.1887	0.0846	0.0685	0.0706	0.0667

Multiplicative model forecasts well especially for first several weeks as displayed in

Figure 17, since it considers neighborhood effects on the prediction. Its forecasting is not as good as the beginning for a long term. Thus, we can believe forecasting for approximate first 15 weeks after the comparison between observations and the forecasted in 2017 with the same model. A detailed forecasting of Beaver Lake in 2018 is given in Appendix B.

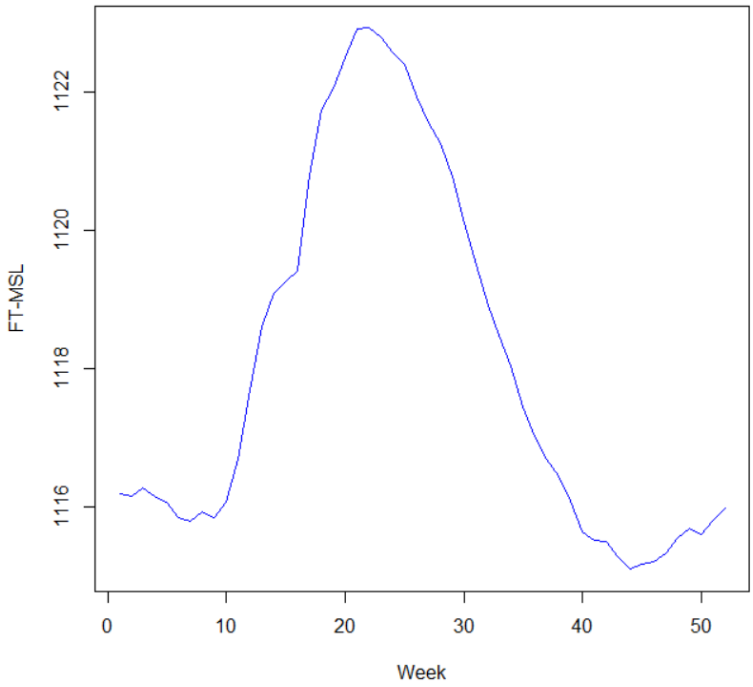

Figure 17: Weekly Prediction for Beaver Lake in 2018

3.4 Models Double Check For Lake Level

Due to the lack of observations in 2018, I fail to make a comparison to examine the correctness of our models for data of lake level. To solve the issues, the forecasted values for 2018 using the same modeling method with lake level data from 1971 to 2016, are obtained, and the graph of comparison is displayed in Figure 18 where the predictions successfully capture the trend but fail to forecast well in the first two months. Besides, the regression method forecasts better for most dates. The reason for this failure is that the lake levels

in the first two months of 2016 are higher than before, which leads to higher forecasting of 2017.

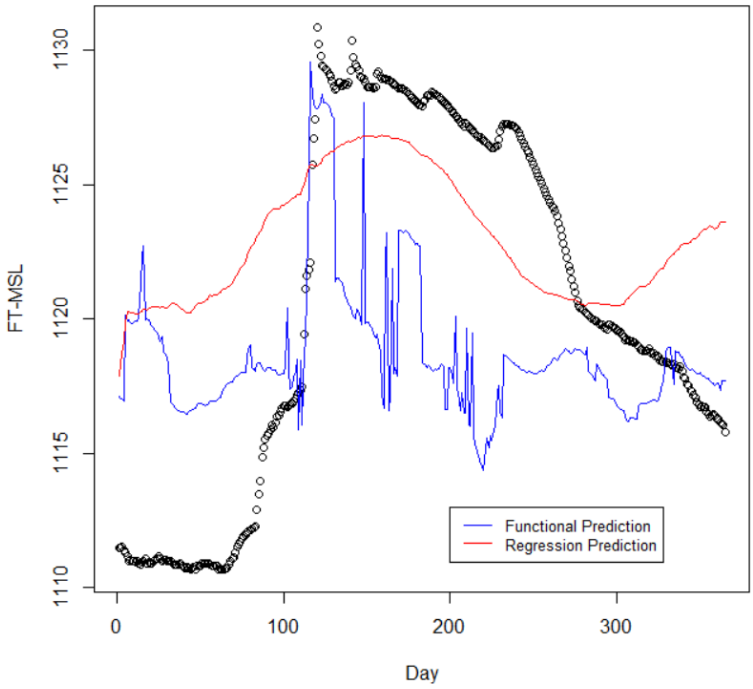

Figure 18: Models Double Check Comparison for Beaver Lake in 2017

4. SUMMARY

This work studies seasonal time series models with application to lake level and weather data. The thesis includes concepts of time series, ARIMA models, three different parameter estimation methods (method of moments, maximum likelihood, and least squares), model diagnostics, and forecasting. Due to seasonality of the weather data and lake level data, three different approaches are proposed to the modeling: regression method, functional ARIMA method and multiplicative seasonal ARIMA method. Regression method considers a polynomial as the trend first, and then takes a model for seasonality. Functional ARIMA method gives an ARIMA model for the time series of any specific date over the past years, and a prediction is given on the same data in 2018 based on the corresponding model. Due to the limit on the period in R, for my data, multiplicative seasonal ARIMA method gives a model on weekly data with period $s = 52$ where the average is taken for per week.

Based on these three different methods, the forecasted values of the year 2018 are compared with observations. They all do well in capturing the trend based on the past values but fail to make correct prediction if something unexpected happen, like outlying observations in the following year, which also explains why they forecast badly in the first two months of lake level in 2017. It's exactly where we need to make improvements in the future. However, the forecasting still makes a difference. If daily forecasting is wanted, regression method is better to be chosen then functional ARIMA method, since regression method considers more in the trend and the other ignores the influence of the neighborhood; if shorter period forecasting is wanted, multiplicative ARIMA method is a better choice owing to higher accuracy for a short term. However, there still need to be more modeling approaches proposed in the future to fit the data and forecast better, so researchers in this field need to continue the efforts and I will be a part of them soon. Functional ARIMA method might be improved by considering the data values of the neighborhood.

REFERENCES

[1] National centers for environmental information, Global Historical Climate Network Daily Data, `https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/`.

[2] Stephanie, Ljung Box test, 2018, `https://www.statisticshowto.datasciencecentral.com/ljung-box-test/`.

[3] Box. G. E. P and Cox. D. R. An analysis of transformations. *Journal of the Royal Statistical Society B*, pages 214-216, 1964.

[4] Franses. P. H. Model selection and seasonality in time series. *Amsterdam: Thesis*, pages 10-15, 1991.

[5] Wei. W. W. S. Time series analysis univariate and multivariate methods. *Boston: Pearson Education*, pages 160-182, 2019.

[6] Cryer. J. D and Chan. K.-sik. Time series analysis: with applications in R. *New York: Springer*, pages 11-245, 2011.

[7] Box. G. E. P. Time series analysis: forecasting and control. *John Wiley & Sons, Inc., Hoboken, New Jersey*, pages 209-213, 2016.

[8] Robinson. P. M. Conditional-sum-of-squares estimation of models for stationary time series with long memory. *Institute of Mathematical Statistics*, pages 130–137, 2006.

[9] Cleveland. W. S. Visualizing data. *Summit, NJ: Hobart Press*, pages 42-171, 1993.

[10] S-Plus 6 for Windows Guide to Statistics, Volume 1, *Insightful Corporation, Seattle, WA*, pages 239-274.

[11] S-Plus 6 for Windows Guide to Statistics, Volume 2, *Insightful Corporation, Seattle, WA*, pages 225-259.

1. APPENDICES

1.1 Appendix A. Forecasting With Regression Method

	01/01	01/02	01/03	01/04	01/05	01/06	01/07
Lake[1]	1116.27	1117.10	1118.577	1120.11	1121.37	1121.847	1121.84
NYCP[2]	-17.67	9.36	19.60	24.21	27.07	28.80	29.84
	01/08	01/09	01/10	01/11	01/12	01/13	01/14
Lake	1121.80	1121.78	1121.78	1121.78	1121.77	1121.76	1121.83
NYCP	30.43	30.76	30.93	31.01	31.05	31.07	31.09
	01/15	01/16	01/17	01/18	01/19	01/20	01/21
Lake	1121.88	1121.93	1121.94	1121.92	1121.99	1122.02	1122.04
NYCP	31.13	31.19	31.27	31.39	31.53	31.71	31.93
	01/22	01/23	01/24	01/25	01/26	01/27	01/28
Lake	1122.07	1122.07	1122.07	1122.06	1122.05	1122.06	1122.06
NYCP	32.18	32.46	32.79	33.15	33.54	33.97	34.44
	01/29	01/30	01/31	02/01	02/02	02/03	02/04
Lake	1122.04	1122.03	1122.08	1122.17	1122.23	1122.22	1122.20
NYCP	34.95	35.49	36.07	36.68	37.33	38.02	38.75
	02/05	02/06	02/07	02/08	02/09	02/10	02/11
Lake	1122.17	1122.11	1122.05	1122.02	1121.99	1121.95	1121.91
NYCP	39.51	40.31	41.14	42.01	42.92	43.86	44.83

[1]Forecasting for lake level of Beaver Lake.
[2]Forecasting for maximum temperature in NYCP.

Appendix B. Weekly Forecasting

	1st[1]	2nd	3rd	4th	5th	6th	7th
Lake	1116.20	1116.15	1116.28	1116.15	1116.08	1115.85	1115.80
NYCP	30.45	42.00	29.88	23.00	44.52	40.12	47.41
	8th	9th	10th	11st	12nd	13rd	14th
Lake	1115.93	1115.85	1116.08	1116.70	1117.70	1118.60	1119.09
NYCP	68.58	64.90	95.44	102.88	107.20	128.35	143.41
	15th	16th	17th	18th	19th	20th	21st
Lake	1119.26	1119.42	1120.80	1121.74	1122.06	1122.51	1122.91
NYCP	166.10	182.34	185.60	201.19	209.62	213.34	231.46
	22nd	23rd	24th	25th	26th	27th	28th
Lake	1122.92	1122.78	1122.57	1122.40	1121.94	1121.56	1121.27
NYCP	253.83	249.20	256.64	272.53	284.31	292.19	290.84
	29th	30th	31st	32nd	33rd	34th	35th
Lake	1120.78	1120.11	1119.50	1118.92	1118.49	1118.03	1117.45
NYCP	303.93	289.25	295.44	287.12	281.56	276.08	277.27
	36th	37th	38th	39th	40th	41st	42nd
Lake	1117.02	1116.70	1116.46	1116.10	1115.65	1115.53	1115.51
NYCP	267.65	248.95	237.53	224.50	209.08	191.75	178.57
	43rd	44th	45th	46th	47th	48th	49th
Lake	1115.29	1115.11	1115.18	1115.21	1115.34	1115.57	1115.70
NYCP	163.00	151.93	134.99	123.93	105.99	102.99	76.00
	50th	51st	52nd				
Lake	1115.62	1115.82	1115.99				
NYCP	68.72	64.95	61.84				

[1]The first week in 2018, and the rest is similar.