UIC University of Illinois at Chicago

Time Series Analysis and Forecasting of Water Levels of Lake Erie

Abhinav Ramesh - 6767803406

Rishika Handa - 664099253

Shivank Kejriwal - 668378710

Parth Koul - 654064575

December 3, 2018

Introduction

A sequence of observed data which is ordered in time, is called a time series.

Time Series analysis comprises of the methods for analyzing the time series data by trying to identify and extract meaningful information from the available raw data.

The goal of such a quantitative analysis of data is to develop a concise but comprehensive characterization of the underlying system in the form of a mathematical model, therefore, this model can be used to analyze the system and predict its behavior under a changing environment. The information obtained from such an analysis can be further employed to alter the possible factors and variables in the system to achieve an optimal performance in some sense. This project is aimed at modelling the water levels of Lake Erie and attempts to forecast lake levels with the help of the model.

Lake Erie

Lake Erie is the fourth-largest lake (by surface area) of the five great lakes in North America, and the eleventh-largest globally if measured in terms of surface area. It is situated on the International Boundary between Canada and United States .The main natural outflow from the lake is via the Niagara River, which provides hydroelectric power to Canada and the U.S. as it spins huge turbines near Niagara Falls at Lewiston, New York and Queenston, Ontario. It is the shallowest of the Great Lakes with an average depth of 10 fathoms 3 feet (63 ft (19 m) and a maximum depth of 35 fathoms (210 ft; 64 m)

Data

Our data has been obtained from a time series data repository which comprises of the levels of the Lake Erie. The data set consists of lake water level above the mean sea level(msl) in meters.

The data set comprises of the levels of Lake Erie over the period of year 1921 - 1970 and is recorded monthly giving a total of 600 data points. Figure(xxx) represents time series data of Lake Erie (mean=14.85, variance= 4.09) with 600 observations of water level from January 1921 – December 1970

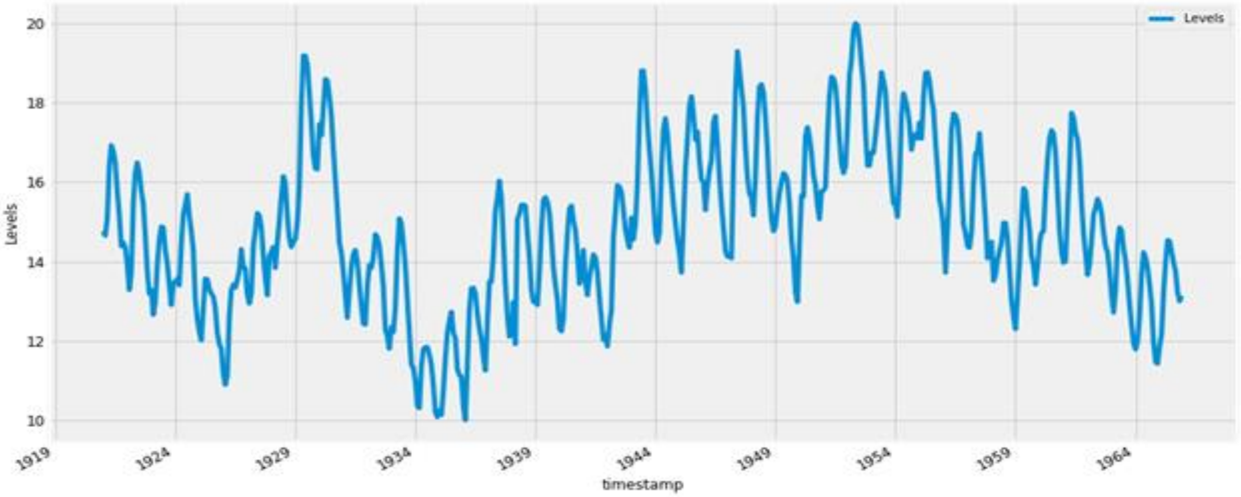


Figure1: Lake Erie Water Level Data

Modeling Procedure

The data set was divided into testing (1 to 540) and validation data points (541 to 600). The study analyzed the deterministic parts of the test data. This was then removed from the raw data to arrive at a stationary stochastic data. The ARMA modeling was performed on this stochastic data. Once all the deterministic and ARMA parameters were identified, a joint optimization algorithm was run to evaluate the initial estimates. Then the model was used to forecast 1 till 5 steps ahead from the last data test data point (540).

Deterministic Trends and Seasonality

The procedure of modeling nonstationary series is developed by considering data with different kinds of trends and seasonality. For the purpose of removing deterministic trend, and for obtaining initial values of the parameters of the final deterministic plus stochastic model, we will first fit the deterministic model alone. In this preliminary fitting of the deterministic part we fit the data in the general non-stationary model, which is represented by the following equation.

$$y_t = \sum_{j=1}^{s} A_j e^{k_j t} + X_t$$

After the equation is fit, the values of A_j and K_j are recorded and further conclusion about he model is made according to the following criteria

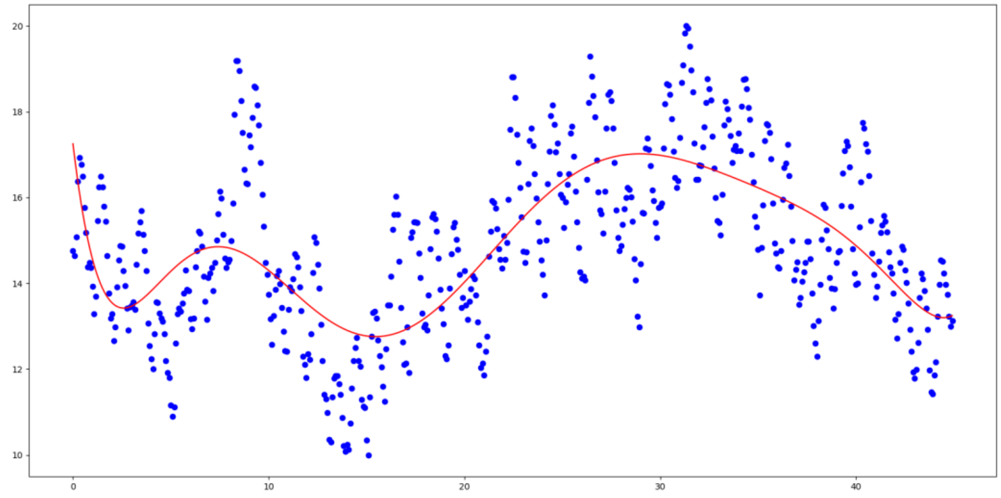
- When all the A_j's are zero, it indicates a zero mean stationary series.
- When only one A_j is non-zero and the corresponding K_j is zero, it indicates a constant mean stationary series.
- When the exponents K_j are real, very small but non-zero, the series has polynomial trends.
- When K_j are real, large and positive or negative, the model has increasing or decreasing exponents trends
- When the K_j are complex conjugates
    - ➢ Negative real part: damped sine/cosine trends
    - ➢ Zero real part: exact sine/cosine with noise
    - ➢ Positive real part: sine/cosine with increasing amplitude

On fitting the model with the above equation, the obtained value of K_j was real, very small but non-zero, which confirmed the presence of polynomial trend in the data. Further, the polynomial trends were run for successive orders increasing from 1 to 9, and the residual sum of squares (RSS) of each increment in the model was compared by conducting the F-test. The conclusion was that there existed an insignificant decrease in RSS from the 8th order to 7th, thus, the polynomial trend attained was of the 8th order. The obtained values of the parameters were as follows.

$$f(x) = (9.26578E - 10) * x^8 - (1.82172E - 10) * x^7 + (1.47307E - 10) * x^6 - (0.000628565 - 10) * x^5 + (0.01506658) * x^4 - (0.1994362) * x^3 + (1.341838) * x^2 - (3.851043) * x + 17.25237$$

(where f(x) is the value of level at time x)

Figure 2 shows the 8th Order fit to the data. It is noticeable that this fit seems similar to the one observed in the decomposed plot.



Levels of Lake Erie, and in general, of water bodies at a similar or higher latitude, are governed by seasonality due to the difference in the lake levels between the lower water levels during winter months versus the higher levels during summer months. To correctly model this periodicity into the time series, the following equation along with the polynomial trend was fit into the data

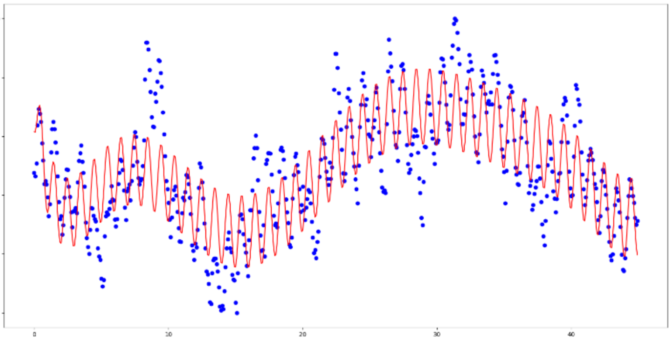
$$f(x) = (9.26578E - 10) * x^8 - (1.82172E - 10) * x^7 + (1.47307E - 10) * x^6 - (0.00062856) * x^5 + (0.01506658) * x^4 - (0.1994362) * x^3 + (1.341838) * x^2 - (3.851043) * x + 17.25237 + B_j e^{b_j x}[C_j \sin(j\omega x) + \sqrt{1 - C^2} cos(j\omega x)]$$

(where f(x) is the level of lake at time x)

The table to the right shows the estimated parameters of this fit. Such a low value of b_j is indicative of a low but existing exponential increase in the amplitude of the periodicity of the data. After conducting F Tests between successive orders of periodicity, it was evident that only one pair of sine-cosine trends were significant in the data. This sort of matches the intuition with the water bodies that have waxing and waning in the summer and winter months.

Coefficients	**Values**
B1	-1.194480335
b1	0.003306536
C1	-0.254442993
x^8	9.564E-10
x^7	-1.87485E-07
x^6	1.51212E-05
x^5	-0.000643728
x^4	0.015400095
x^3	-0.203569291
x^2	1.368711614
x^1	-3.927646042
x^0	17.31027052

Figure 3 shows the fit of the above data to the raw data. This fit "covers" more data points than the previous figure.

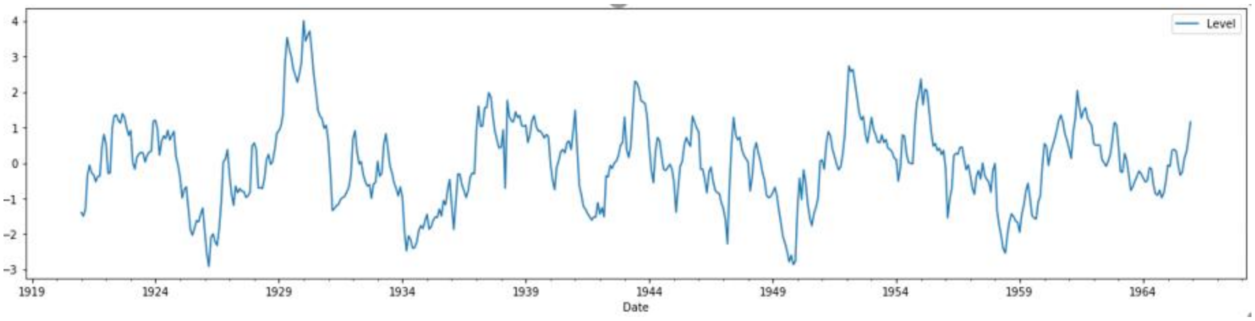


The 8th order polynomial along with the periodicity encapsulates the deterministic part of the data. The ARMA modeling is performed on a stationary dataset, i.e., a data set which is bereft of any deterministic factors. The following equation summarizes this relation

$$y_{raw} = y_{deterministic} + y_{stochastic}$$

Thus, the deterministic part was removed from the raw dataset to obtain the stochastic (time origin independent) data for ARMA modeling.

Figure 4 shows the plot of stationary data point the over the range of time.



The plot above shows the random fluctuations in the lake levels, which, if understood properly could lead to excellent modeling of the levels.

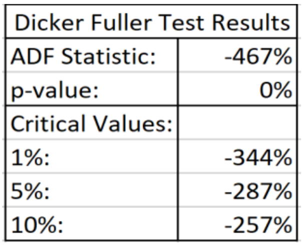
Checks for stationarity:

- Look at Plots: Time series plot of data and visually check if there are any obvious trends or seasonality.
- Summary Statistics: The summary statistics for your data for seasons or random partitions and check for obvious or significant differences.
- Statistical Tests: Statistical tests to check if the expectations of stationarity are met or have been violated.

Stationarity of the obtained residuals from the raw data was checked by performing Dicker Fuller Test.

Dicker Fuller Test Results:

- Null Hypothesis (H0): If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure.
- Alternate Hypothesis (H1): The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have time-dependent structure.



Dicker Fuller Test Results	
ADF Statistic:	-467%
p-value:	0%
Critical Values:	
1%:	-344%
5%:	-287%
10%:	-257%

Results: p-value <= 0.05: Reject the null hypothesis (H0), the data does not have a unit root and is stationary.

ARMA Modeling

In the statistical analysis of time series, autoregressive–moving-average (ARMA) models provide a parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials, one for the autoregression (AR) and the second for the moving average (MA).

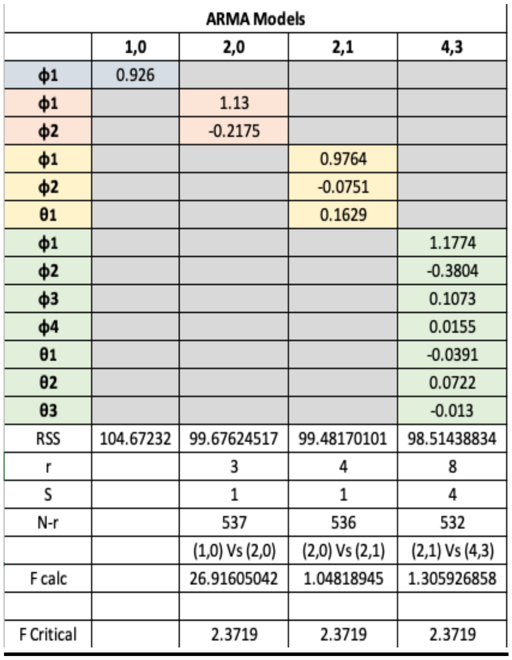
In time series of data, the ARMA model is a tool for understanding and, perhaps, predicting future values in this series. The AR part involves regressing the variable on its own lagged (i.e., past) values. The MA part involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past. The model is usually referred to as the ARMA (*p*, *q*) model where *p* is the order of the AR part and *q* is the order of the MA part.

A mathematical model is only a good approximation to the actual system behavior; therefore, it is possible to approximate closely a stable stationary stochastic system by an ARMA (n, n-1) model for sufficiently large n.

Modeling:

The ARMA (n, n-1) strategy in trying to reduce a dependent series of data into an independent one. The unconditional regression or the dynamic aspect of ARMA (n,n-1) models show that increasingly complex dynamics can be represented with increasing n by the Green's function or the autocovariance function. Such a representation is again equivalent to the ARMA (n, n-1) model, and it can be shown that an arbitrary set of data can be represented by such a model with large enough n. It is advantageous to increase n in steps of two and fit a sequence of ARMA (2n, 2n-1) models until the reduction in the sum of squares of 's is insignificant as judged by the F-criterion.

The stationary data of Lake Erie was using ARMA (2n, 2n-1) approach. ARMA (2,1) and ARMA (4,3) was fitted on the data, and the reduction in RSS was judged by the F-test criterion. It was found to be insignificant. So, the data was fitted by AR (2) model and F-test was performed. Increase in RSS was significant hence AR (1) model was fitted on the data. Comparing the RSS of AR (1) and AR (2), the increase was not significant. Hence, it was concluded that AR(2) model best fits the data. Table (a) shows the results of F-test.



ARMA Models				
	1,0	2,0	2,1	4,3
ϕ1	0.926			
ϕ1		1.13		
ϕ2		-0.2175		
ϕ1			0.9764	
ϕ2			-0.0751	
θ1			0.1629	
ϕ1				1.1774
ϕ2				-0.3804
ϕ3				0.1073
ϕ4				0.0155
θ1				-0.0391
θ2				0.0722
θ3				-0.013
RSS	104.67232	99.67624517	99.48170101	98.51438834
r		3	4	8
S		1	1	4
N-r		537	536	532
		(1,0) Vs (2,0)	(2,0) Vs (2,1)	(2,1) Vs (4,3)
F calc		26.91605042	1.04818945	1.305926858
F Critical		2.3719	2.3719	2.3719

Table(a): ARMA Modeling Parameters

From AR(2) model result table $\lambda_1 = 0.8839$ and $\lambda_2 = 0.2361$, hence the model fitted is asymptotically stable. Also due to the absence of MA parameter in the model, there is invertibility restrictions, (i.e. Always bounded). Green's function can be given by

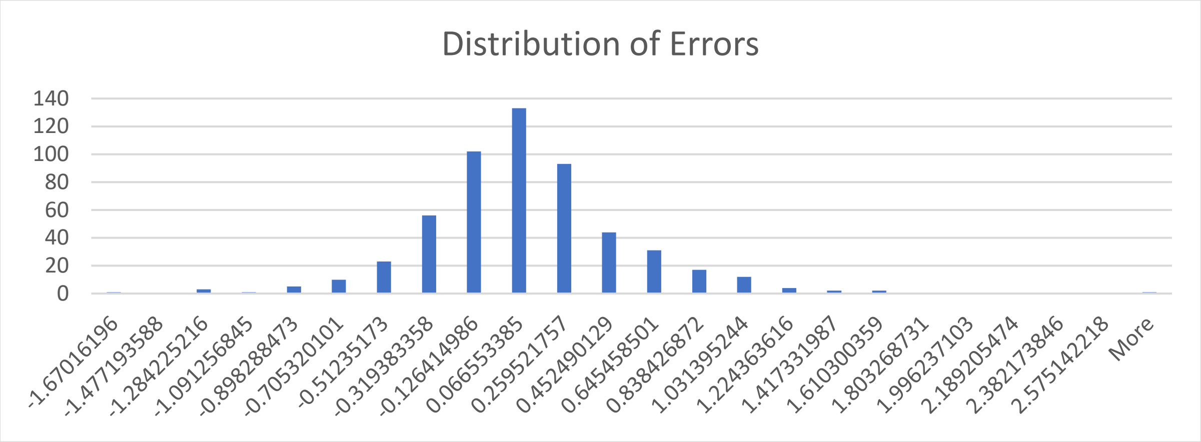
$$G_j = (1.385)(0.8839)^j + (0.3858)(0.2461)^j$$

Model Performance

Below table depicts all the parameters estimated thus far

	Final AR+Trend Model
B1	-1.194480335
b1	0.003306536
C1	-0.254442993
x^8	9.564E-10
x^7	-1.87485E-07
x^6	1.51212E-05
x^5	-0.000643728
x^4	0.015400095
x^3	-0.203569291
x^2	1.368711614
x^1	-3.927646042
x^0	17.31027052
ϕ1	1.1283
ϕ2	-0.2161

Figure 5 depicts is the distribution of errors



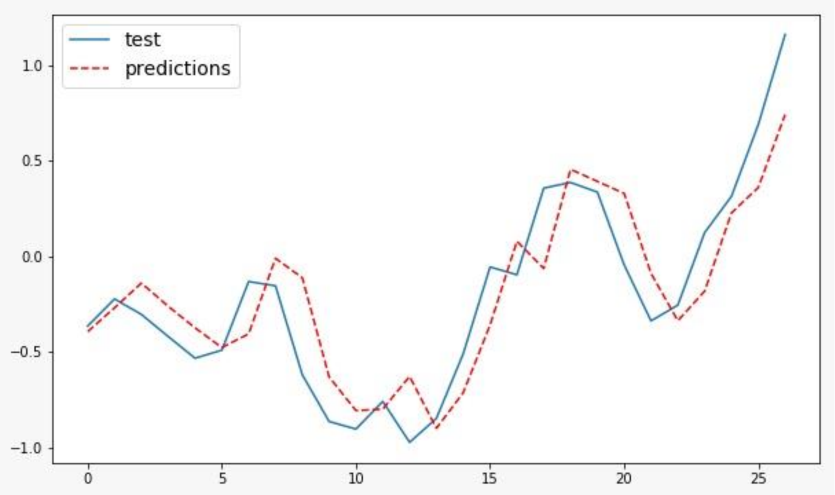
Since the distribution is Normal and the average of residuals is close to mean, we can conclude that the errors is Normally Independently Distributed with a mean almost equaling zero (0.0016) and having a standard deviation of 0.434

Thus, the model can be concluded to be adequate. After estimating the values of the coefficients of AR parameters, the RSS, RMSE and R^2 values of the entire model was calculated as follows:

RSS TS+AR	101.6203370880
RMSE	0.188185809
R Squared	95.40%

It can be noted that the model that was fit captured 95.4% of the variance of the original data which is indicative of a good fit.

The In-Sample forecast results were analyzed as another means of validating the model. The forecasts were satisfying as evident from the below plots and an RMSE value of 0.242



Joint Optimization of Parameters

One of the pitfalls of optimization is that in the search grid, the operation might yield a local extremum. To ensure that the optimization algorithm reaches global extremum, the search must be provided initial decent estimates. After the estimation of all parameters by the procedure described above, a joint optimization algorithm can be run to estimate all the parameters together. With the previously estimated parameters as good starting estimates, a global extremum could be reached. Below table depicts the Joint Optimized parameters alongside the previous estimates.

	Final AR+Trend Model	**Joint optimised Model**
B1	-1.194480335	-1.239428209
b1	0.003306536	0.001610872
C1	-0.254442993	-0.255465514
x^8	9.564E-10	9.79E-10
x^7	-1.87485E-07	-1.91E-07
x^6	1.51212E-05	1.54E-05
x^5	-0.000643728	-0.000653223
x^4	0.015400095	0.015591157
x^3	-0.203569291	-0.205690484
x^2	1.368711614	1.380680099
x^1	-3.927646042	-3.955713264
x^0	17.31027052	17.32593332
ϕ1	1.1283	1.126068945
ϕ2	-0.2161	-0.213593268
RSS	101.6203371	101.2721427

What's noticeable is that the joint optimization yields an RSS value that is very similar to the initial estimated model. This indicates that the deterministic trends and seasonality were correctly captured and the initial modeling was robust.

Fig xxxx depicts how well the Joint Optimised parameter fits the raw data. Note the apparent difference in the fits of initial versus the joint optimized model.

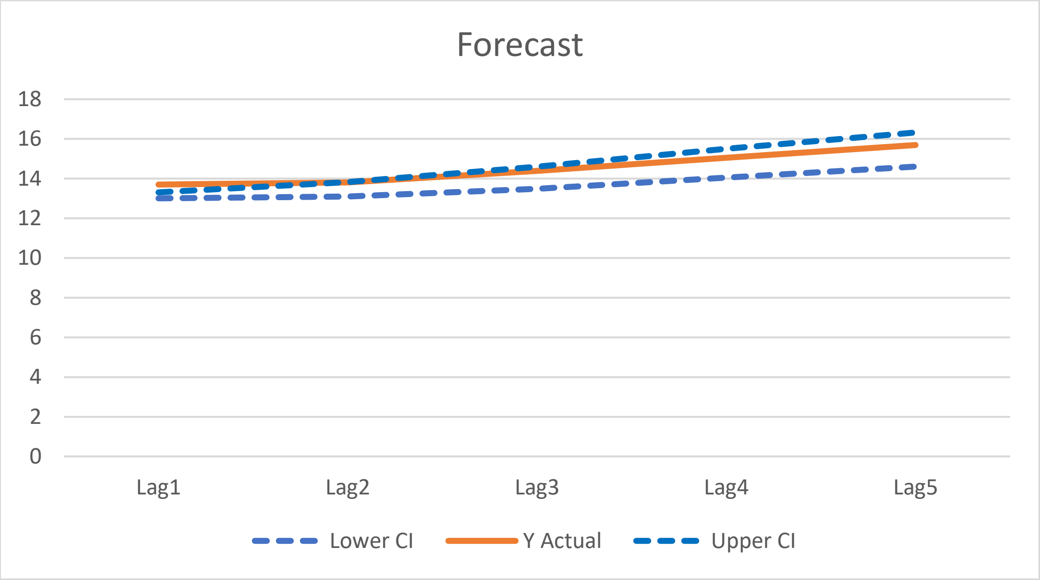
Forecasting

The principal aims of time series modelling is prediction or forecasting. It is called extrapolation since it involves extrapolating the value steps ahead from the knowledge of the series and its structure. The main reason to undertake the analysis of a system from its observed data is to predict, forecast or to extrapolate its behavior at future times. The manipulation or control of the behavior is also based on prediction, so prediction is also viewed as the heart of control and regulation. In this project, after obtaining the joint optimization estimates, 5 data points were calculated after the end of the training data. The method used in this analysis to forecast values some steps ahead of the last known value is called the Forecasting by Conditional Expectation. Below is the summary table of the forecast values along with the residuals and square of residuals.

	Lower CI of Predicted Value	Y Actual	Upper CI of Predicted Value	Point Forecasted Value	Residual Squared
Lag1	13.00466	13.7	13.30837624	13.15551181	0.296467385
Lag2	13.09377	13.814	13.81858896	13.44224221	0.138203853
Lag3	13.4829	14.383	14.60309663	13.98504507	0.158368128
Lag4	14.05235	15.047	15.50146582	14.63850753	0.166866102
Lag5	14.60444	15.693	16.32308293	15.21424889	0.229202625
				RMSE	0.197821619

It can be noted that the actual lake level values of all the lags, except the first one, lie within the predicted confidence interval of 95%. It can also be noted from the last column that as one forecasts farther into the future, the error increases (which is intuitive). This is overcome by updating the forecast at every step of forecasting.

Figure 6 shows the forecasted values along with the confidence interval plots.


Forecast
Lag1
Lag2
Lag3
Lag4
Lag5
Lower CI
Y Actual
Upper CI