**An Analysis of Reservoir Levels in California (WIP)**

*Jeremy Chan, Fourth-Year Economics and Statistics Major*

**Table of Contents**

**1 Summary 2**

**2 Introduction 2**

**3 Methods 3**

3.1 Initial Data Cleaning and Inspection 3

3.2 Fitting a Time Series Model 3

3.3 Forecasting 5

**4 Results and conclusions 6**

4.1 Initial Data Cleaning and Inspection Results 6

* 1. Fitting a Time Series Model Results 7

4.3 Forecasting Results 9

**5 Code Appendix**  12

**1 Summary**

This report contains an analysis of water reservoir levels in 11 selected California reservoirs. The ultimate goal of the report was to provide a summary of the current reservoir storage as a percentage of total capacity and to provide a forecast of these values for the year of 2016. As the report was written in May of 2016, there are already some data points from 2016 which will be compared with the forecasted points. To accomplish the task, first the data was plotted and inspected. A stationary time series was then modeled and predictions made from the residuals of that model. Lastly, the forecasts were plotted and listed in the results section. Additionally, the model for each reservoir was used to predict 2014 and compared against actual values to further test the accuracy of the models.

**2 Introduction**

Since 2012, California has faced severe drought conditions, the likes of which the state has not seen since the 1970s. As not only the most populous state but also the state which produces the most food[[1]](#footnote-1), California faces unique challenges during a drought. It is, therefore, of great interest to track the level of drought which California faces and to make predictions on if and when the state will return to normal water levels.

According to the United States Geological Service, “a drought is a period of drier-than-normal conditions that results in water-related problems” resulting from “less than normal [rainfall] for several weeks, months, or years.”[[2]](#footnote-2) One key indicator of drought conditions is reservoir levels. During period of drought, reservoirs will The California Department of Water, provides reservoir information on 199 different reservoirs all throughout the state[[3]](#footnote-3). Of these, 199 reservoirs, eleven were selected for analysis in this report. The 11 that were selected were part of a group of 12 which the California Department of Water uses for their daily report on the conditions for “major reservoirs.”[[4]](#footnote-4) These 11 reservoirs create a fairly representative picture of the overall reservoir levels in California and have been a steady indicator of drought levels. Thus, it is of interest to analyze and predict the levels in these reservoirs.

All of the data from the remaining reservoirs was retrieved through a query function from the California Department of Water Resources through the California Data Exchange Center. The data for each reservoir’s capacity was queried at a monthly level, from the first recorded measurement to the most recent measurement of April 2016 and measured in acre-feet, with one acre-foot being equivalent to 325,851 gallons of water.

The ultimate goal of this report is to provide an overview of the selected reservoirs’ raw storage as a percentage of its capacity, provide a forecast of these values for the 2016 calendar year for each reservoir, and to compare each reservoir’s forecasted storage levels with the actual observations from 2016, up through the present time. To accomplish this, the first step is to retrieve, inspect, and clean the data which has been retrieved. Any outliers will be removed if they are deemed as incorrect readings and any non-stabilized variance will be corrected for. Additionally, if there are any reservoir readings which appear clearly non-linear, they will be adjusted for on a case-by-case basis.

The next step is to remove any deterministic component from each time series. Trend and seasonality will be removed and once stationarity has been established, a model will be fitted to the remaining residuals. From that model, a 12-month forecast for the calendar year of 2016 will be created. If the residuals from the model are normal, then a prediction interval will be provided in addition to the point forecasts.

Finally, the model will be evaluated against the actual values from 2016. Since 2016 is an El Niño year, the resultant models may not be accurate. As such, each model will also be used to forecast 2014 and compared against actual 2014 values to determine if the models are accurate for both El Niño and non-El Niño years.

**3 Methods**

In this section, the methods used for analysis is detailed. The analysis will be split into the following sections: Data Retrieval, Data Cleaning, Data Visualization, Removing Deterministic Components, Fitting a Time Series Model, and Forecasting. Due to the number of reservoirs that are being analyzed, all of the above tasks were combined into a single function in R which simplifies the actual coding analysis aspect of this report.

**3.1 Data Retrieval**

All of the data contained in this report was retrieved from the California Department of Water through the California Data Exchange Center (CDEC). To automate the retrieval of this data, the package “sharpshootR” was used. The package contains a function called CDECquery which retrieves the reservoir levels for a given reservoir between a certain date range and at a certain frequency.

For this report, the following 11 reservoirs were used as a representative sample of California’s overall reservoir levels (Reservoir abbreviations denoted in parentheses): Trinity Lake (CLE), Lake Shasta (SHA), Lake Oroville (ORO), New Melones (NML), Folsom Lake (FOL), Don Pedro (DNP), San Luis (SNL), Millerton Lake (MIL), Pine Flat (PNF), Castaic Lake (CAS), and Lake Perris (PRR).

The data was retrieved on a monthly basis spanning from when each reservoir was first opened up through April 2016. Each reservoir’s data contained the following: date and time of each observation, year of each observation, month of each observation, the raw reservoir storage, and the ID of each reservoir (the three letter reservoir abbreviation).

After retrieving the raw storage readings for each of these reservoirs, the capacity of all California reservoirs was retrieved from the CDEC[[5]](#footnote-5). From this table of reservoir capacities, the specific capacities for our reservoirs of interest were saved within R.

The raw reservoir storage was divided by each reservoir’s overall capacity, resulting in a percentage of capacity for each reservoir (henceforth referred to as a reservoir’s capacity level). Capacity level was used as the primary unit of observation for ease of interpretation. Since converting to percentage of capacity is a linear transformation, it should not affect the final results of the time series model and forecast.

Note: Each reservoir was stored in a list; moving forward, an individual item in this list will be referred to as a “reservoir object” which contains all of the data for a single reservoir including date and time of each observation, year of each observation, month of each observation, the raw reservoir storage, the ID of each reservoir (the three letter reservoir abbreviation), and the reservoir’s capacity level.

**3.2 Data Inspection**

After the data had been retrieved, each reservoir’s historic capacity levels were plotted to look for the following problems: inconstant variance, large number of NA values, any sharp trend changes in the data, and outliers in the observations. Following that step, notes were made on which reservoirs needed additional inspection based on immediate visual inspection.

If inconstant variance was found, a log transformation would have been introduced to stabilize the variance so that the time series could be properly modeled. However, no times series were found to have seriously inconstant variance.

To check for NA values, each reservoir had its total number of NA values summed up. If they had a NA values that were spaced at least two values away from another NA value, the value was handled by imputation (detailed in Section 3.4).

Sharp trend changes and outliers in the data were handled on a case-by-case basis. To address a sharp trend change, historical data was consulted to see if the change in trend was due to a statewide drought or water shortage. Additionally, that same time period was cross-checked with other reservoirs to see if there was a trend across the state. All sharp trend changes and outliers were explained by either historical water occurrences or documented changes to the reservoir, thus, no further transformation was required.

**3.3 Data Cleaning**

To clean the data, the aforementioned NA checking process was utilized. NAs which were found to be far enough away from each other were handled by imputation. Any NA values which did not fit this process were handled on a case-by-case basis and are described in the individual results (detailed in Section 4). Additionally, as there was only reservoir with a major drop not related to state-wide water conditions, the data for that reservoir was handled in a singular manner (detailed in its individual results in Section 4).

**3.4 Forecasting**

To forecast the 2016 calendar year for each reservoir, a generalized function was created which would take a reservoir object in as an input. The function (henceforth referred to as forecast.all) then works to derive a stationary time series and model that time series. Forecast.all was written in such a way that it will predict the calendar year for the latest month it has data for, e.g., if your latest observation is January 2017, it will predict all of 2017. For the purposes of this project, 2016 was the predicted year.

Forecast.all first generates a time series of reservoir capacity percentages based on the first January observation and the last December observation in preparation for using sum of harmonics to remove seasonality from the time series (this abbreviated series will be henceforth referred to as “year series”). After creating the time series, any missing values are imputed by calculating an average of the two previous and two following observations around the NA value. If there are less than two values before after the missing value, a simple average of the previous and next value is used.

Next, forecast.all removes any deterministic trends from the year series. First, a first-order difference operation is used to remove any trend. Next, a sum of harmonics operation is used to remove the seasonality from the year series and fit a model. The residuals are then obtained from the de-trended and de-seasonalized year series.

These de-trended residuals are then evaluated for stationarity using both the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test and the Augmented Dickey-Fuller (ADF) tests for stationarity. If the de-trended residuals are found to be not stationary by either of these tests, a warning is produced by forecast.all stating that the resultant series is not stationary, but the function continues the process of forecasting.

Next, the de-trended residuals are tested for independence using the Ljung-Box test for independence. If the de-trended residuals are found to be dependent, forecast.all produces a warning, indicating that other methods of de-trending the year series may be necessary.

After testing for independence, the de-trended residuals are tested for normality for the use of creating a prediction interval. The residuals are tested using the Shapiro-Wilk test and forecast.all will produce a warning if the residuals are found to be non-normal by this test. Additionally, a histogram of the residuals is produced, which is often a better indication of normality of the residuals. The histogram is displayed within the R plot panel. If the residuals are not normal, the prediction intervals may not be accurate.

Finally, the forecasting begins with forecasting the noise from the de-trended residual model. 12 periods corresponding to the 12 months of 2016 are forecasted at a 95% confidence level. The seasonal values are then fitted back onto the noise forecast, along with the upper and lower bound of the noise forecast, creating a 95% prediction interval. Finally, the seasonalized noise forecast and upper and lower seasonalized bounds are undifferenced, creating three vectors: a point forecast, an upper bound, and a lower bound.

Forecast.all then generates four separate time series: one with all of the observed values from the CDEC, one with the point forecast for 2016, one with the upper prediction bound for 2016, and one with the lower prediction bound for 2016. Finally, forecast.all produces a series of plots: the raw data, the cleaned data, the cleaned data with the forecast and prediction interval, and a zoomed-in view of the cleaned data, forecast, and prediction interval including the year predicted and the year prior.

The function thus returns four plots of the data, a histogram of the residuals, and a number of other objects (the fitted model, point forecast, lower prediction bound, upper prediction bound, and residuals) in text form to the user for analysis and presentation.

**3.5 Forecast Comparison**

As a method of analyzing the relative prediction power of the fitted models, a separate forecasting function, forecast.2014 was created. Forecast.2014 is a replication of the forecast.all function, but it specifically creates a forecast for 2014, as 2014 was the last year without an El Niño and without any effects of the onset of an El Niño as 2015 had. This serves as a visual check to assess each reservoir model’s predictive during non-El Niño years in comparison to its predictive power during El Niñoyears. After running each reservoir’s data through forecast.2014, the graphs were inspected in comparison to the 2016 predictions.

**4 Results**

The results section will be divided by individual reservoir for ease of finding specific reservoir forecasts. Please refer to the table of contents for a page numbers for each reservoir.

**4.1 Trinity Lake Reservoir (CLE) Forecast Results**

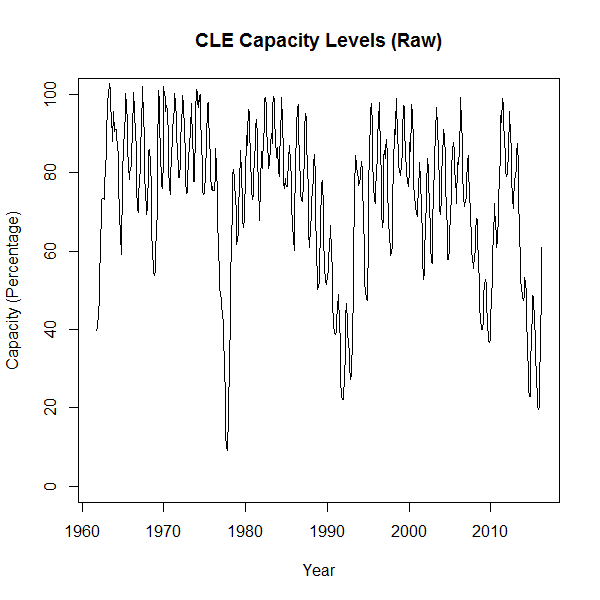
**4.1.1 Trinity Lake Reservoir (CLE) Data Retrieval**

Trinity Lake (CLE) Reservoir data was retrieved from the CDEC using sharpshootR for the date range of October 1961 to April 2016 at a monthly level, totaling 655 observations. The raw reservoir storage was then divided by Trinity Lake’s capacity of 2,447,650 acre-feet of storage capacity to find the capacity percentage of every observation.

**4.1.2 Trinity Lake Reservoir (CLE) Data Inspection**

Upon initial inspection, the data for Trinity Lake (shown below in **Figure 1**) did not appear to require any additional transformation or cleaning.

**Figure 1.** Raw data for Trinity Lake Reservoir (CLE)



After checking for any NA values, none were found, meaning that no imputation was required for this reservoir. Sharp drops in the late 1970s, early 1990s, and early 2010s were accounted for by statewide drought conditions during those times. As such, the observations are recorded as intended and are not incorrect measurements.

**4.1.3 Data Cleaning**

No data cleaning was required for Trinity Lake capacity percentage observations.

**4.1.4 Forecasting**

1. http://www.ers.usda.gov/faqs.aspx [↑](#footnote-ref-1)
2. http://ca.water.usgs.gov/data/drought/ [↑](#footnote-ref-2)
3. http://cdec.water.ca.gov/misc/resinfo.html [↑](#footnote-ref-3)
4. http://cdec.water.ca.gov/cdecapp/resapp/getResGraphsMain.action [↑](#footnote-ref-4)
5. http://cdec.water.ca.gov/misc/resinfo.html [↑](#footnote-ref-5)