Time Series Analysis of Murray Darling Basin River System

ABSTRACT

The Murray-Darling Basin (MDB) has an enormous area in the southeast region of Australia, which supports around 3 million of the human population and is home to around 35 endangered species. MDB basin has two major rivers that are River Darling and the Murray River (Murray Darling Basin Authority [MDBA], 2022e). The Murray River provides approximately 60% of the water to Adelaide city (Department for Environment and Water [DEW], 2022). More notably, European settlements in the past have led to an increase in salinity, blue-green algae, droughts and acid sulfate in soils (MDBA, 2022f). Due to this, there is a requirement to forecast the Salinity, Water Level and Water Temperature of the River Murray to prevent any disastrous situation in future. This research focuses on finding accurate and reliable models to accomplish the forecast of the above three variables. More notably, this study is divided into two parts: (1) In the first part of the research project paper, two forecasting models, ARIMA and SARIMA models, were utilized to forecast Salinity, Water Level and Water Temperature at four locations, specifically Biggara, Albury, Colignan and Murray Bridge. It was found that for all these four locations, the SARIMA model was better than the ARIMA model for Water Level and Water Temperature because the differenced values of MAPE (MAPE of ARIMA minus MAPE of SARIMA) were greater than that of the differenced values of RMSE (RMSE of ARIMA minus RMSE of SARIMA). Furthermore, it was observed that for Salinity at Albury and Biggara, ARIMA was the best model as its RMSE, and MAPE values were less than that of the SARIMA model. Similarly, for Salinity at Colignan and Murray Bridge, SARIMA was the best model as its RMSE, and MAPE values were less than that of the ARIMA model. (2) In the second part of the research project paper, the RBF SVM model's forecast performance was compared with ARIMA and SARIMA to forecast the Salinity of the same four locations from above. More importantly, because the performance of SARIMA was the best for Water Level and Water Temperature, these variables were used as predictor variables for the response variable Salinity. It was discovered that for all four locations, SVM RBF was undoubtedly the best model since it had lower RMSE and MAPE values when compared to both ARIMA and SARIMA models. To conclude, according to our research, for Water Level along with Water Temperature SARIMA model should be preferred, and for Salinity, RBF SVM should be favoured for the four locations mentioned above.

1. INTRODUCTION

The Murray-Darling Basin (MDB) is an area made up of several intertwined lakes and rivers, out of which two main rivers are the River Darling and the Murray River. Furthermore, it provides shelter to 35 threatened species along with 120 waterbird species. More importantly, a human population of approximately 2 million resides in the MDB Basin. Also, MDB Basin generates a revenue of about $11 billion annually due to tourism. A major chunk of agricultural produce comes from the MDB Basin, which includes rice, grapes and dairy (Murray Darling Basin Authority [MDBA], 2022e).

Due to European settlements in the early 19^{th} century in the MDB Basin, the industries and people grew in number, which led to an increase in water consumption from the MDB River System. This abnormal use of water caused various calamities such as Drought (due to a decrease in Water Levels and River Water Flow along with an increase in Salinity, Water and Air Temperature), High Salinity, Blue-Green Algae and Acid Sulfate Soils. To tackle these issues, the Australian Parliament passed the Water Act in 2007, which made the Murray Darling Basin Authority (MDBA) in charge of maintaining the MDB Basin and making it a healthier working Basin. Moreover, a Basin Plan was also implemented so that everyone can share water in a sustainable manner (MDBA, 2022f).

Therefore, there is a requirement to do a time series analysis of the Murray Darling Basin so that we can observe whether the Water Act and the Basin Plan have improved the condition of water in terms of Salinity, Water Level and Water Temperature. More notably, the time series of Salinity, Water Level and Water Temperature needs to be forecasted to prevent any disastrous situation that is going to occur in future. For this, we would have to find a good model for predicting future values of the above variables.

2. BACKGROUND (PART-A)

Musarat et al. (2021) conducted a study where they modelled the streamflow of the Kabul River with the help of a statistical tool called Auto-Regressive Integrated Moving Average (ARIMA). It was observed that the ARIMA model with the lowest value of the Akaike Information Criterion (AIC) had the capability to explain about 92% of the dataset for forecasting. Furthermore, the study concluded that the streamflow would maintain a lower bound of 10 cubic metres per second and an upper bound of 250 cubic metres per second till the year 2030 (Musarat et al., 2021).

Chen et al. (2018) examined the average monthly temperature from 1951 to 2017 in Nanjing, China, with the help of the Seasonal Auto-Regressive Moving Average (SARIMA) Model. The researchers selected data from 1951 to 2014 as the training set, whereas data from 2015 to 2017 was chosen as the testing dataset (Chen et al. 2018). Chen et al. (2018) described the best SARIMA model as having the least AIC value. Furthermore, it was observed that SARIMA Model gave a Mean Squared Error (MSE) of 0.84, 0.89 and 0.94 for the years 2015, 2016 and 2017, respectively, which indicates good forecasting accuracy (Chen et al. 2018).

Therefore, we will utilize ARIMA and SARIMA models to forecast the MDB Basin's Salinity, Water Level, and Water Temperature. Furthermore, we will compare their forecast performance, and we will select the best model for each variable mentioned above in the first part of the research. In addition, we will select the best ARIMA or SARIMA model for each variable based on AIC, Bayesian Information Criterion (BIC) and Corrected Akaike Information Criterion (AICc) and will compare ARIMA and SARIMA models based on Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) values.

3. RESEARCH QUESTIONS (PART-A)

The Research Questions (for the first part of the research) for the Time Series Analysis of the Murray Darling Basin (MDB) River System are:

1) Which model, ARIMA or SARIMA, is better for performing a Time-Series Analysis (Forecasting) for Salinity?
2) Which model, ARIMA or SARIMA, is better for performing a Time-Series Analysis (Forecasting) for Water Level?
3) Which model, ARIMA or SARIMA, is better for performing a Time-Series Analysis (Forecasting) for Water Temperature?

4. DATA

The available data which would help in the time series analysis of the Murray Darling Basin (MDB) River System is very vast. Thus, it needs to be explained, filtered, and cleaned. Therefore, to accomplish this, the Data Section is divided into Available Data, Data Explanation, Data Parameter Selection, Data Location Selection, and Data Cleaning. All these Sub-Sections of Data are elaborated below:

4.1. Available Data and Explanation

The Data which would be utilized in the Time-Series Analysis of the MDB River System is provided by the Murray Darling Basin Authority (MDBA). The Data is retrieved from a web of hydrometric observation sites present at different locations across the MDB River System, which are maintained and operated by the MDBA Authority with the help of state governments. The Data, which includes calculated flow rates, storage and river levels, various water quality attributes and rainfall, is disseminated and recorded with the help of telemetry (Murray Darling Basin Authority [MDBA], 2022a).

The website provided by MDBA Authority for Murray River contains data of key monitoring stations for the public to view and access. The data can be downloaded in Comma Separated Values (CSV) format. The CSV file contains both historic and near-instantaneous (or near-real-time) data, which can be utilized for Time Series Analysis. The data that is near-real-time is documented on a six-hourly basis. Furthermore, every time-series data has a related date and timestamp with it (MDBA, 2022a).

4.2. Data Parameter and Location Selection

The MDBA Authority Data, as explained above, has three key data types or variables, which are explained in detail in Appendix-D under the heading “Data Parameter and Location Selection”.

One of the major concerns of the MDBA Authority is that the Salinity of the MDB River System has increased drastically because the groundwater under the southern part of the MDB River has a high concentration of salt (MDBA, 2022b). Another concern of the MDBA Authority is the concentration of Acid Sulphates in the river water, which are caused due to lower river water levels. The water near the edges of rivers and lakes can become so acidic that it might start to corrode steel (MDBA, 2022c). Lastly, another concern of the MDBA Authority is that the increase in water temperature due to climatic conditions can lead to droughts (MDBA, 2022d). Therefore, the data variables that we chose for time series analysis were Salinity, Water Levels and Water Temperature.

Out of 40 monitoring stations, 4 sites, namely Biggara, Albury, Colignan and Murray Bridge, were selected for this research (MDBA, 2022e). The Biggara monitoring site was chosen because it is the starting point of the Murray River. Whereas Murray Bridge was selected because it is the point where the Murray River ends. Furthermore, Albury was selected because it is at the border of New South Wales and Victoria. While the Colignan was chosen as it is nearer to the border of South Australia and Victoria. To conclude, we chose one ending and starting station with two intermediate sites. These sites are shown in Figure 1.

Figure 1: The four selected locations for Time Series Analysis

4.3. Data Filtering and Imputing

The Data Provided by the MDBA Authority for four locations contained data approximately ranging from the year 1900 to 2022. However, the issue was that the time series data before 2009 was very sparse as it had a lot of missing values. Moreover, as we only wanted to analyse data from the last 12 to 14 years. Therefore, we chose data from 2009 to 2022.

Another task was to impute the missing values present in the time series data from 2009 to 2022. For this Robust Linear Regression (RLR) Model was utilized because it is immune to outliers (John & Sanford, 2013). The values of the time series are fed into the RLR Model, which produces a linear curve with the help of M-estimators. The missing values are replaced with the corresponding values on the linear curve. The independent variable is the time stamp, whereas Water Salinity, Water Level or Water Temperature are taken as dependent variables.

The “M” in M-estimators stands for maximum likelihood function, and M-estimators are of three types i.e., the least square estimator, the Huber estimator, and the Tukey bi-square estimator. The least-square estimator performs badly if the distribution of error or residual values is not normally distributed. Therefore, a combination of Huber and Tukey Bi-square estimator is utilized in RLR Model to overcome this issue (John & Sanford, 2013). The objective and weight functions of the least square estimator, the Huber

estimator, and the Tukey bi-square estimator are shown in Figure 2, here, "k" is the tuning constant, and e is the error term or residual (John & Sanford, 2013).

The data imputation was accomplished with the help of the impute_rlm function in the simputation Library of R Programming Language (Mark, 2022).

Method	*Objective Function*	*Weight Function*
Least-Squares	$\rho_{\mathrm{LS}}(e) = e^2$	$w_{\mathrm{LS}}(e) = 1$
Huber	$\rho_{\mathrm{H}}(e) = \begin{cases} \frac{1}{2}e^2 & \text{for } \lvert e\rvert \le k \\ k\lvert e\rvert - \frac{1}{2}k^2 & \text{for } \lvert e\rvert > k \end{cases}$	$w_{\mathrm{H}}(e) = \begin{cases} 1 & \text{for } \lvert e\rvert \le k \\ k/\lvert e\rvert & \text{for } \lvert e\rvert > k \end{cases}$
Bisquare	$\rho_{\mathrm{B}}(e) = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{e}{k}\right)^2\right]^3\right\} & \text{for } \lvert e\rvert \le k \\ k^2/6 & \text{for } \lvert e\rvert > k \end{cases}$	$w_{\mathrm{B}}(e) = \begin{cases} \left[1 - \left(\frac{e}{k}\right)^2\right]^2 & \text{for } \lvert e\rvert \le k \\ 0 & \text{for } \lvert e\rvert > k \end{cases}$

Figure 2: The Objective and Weight Functions of different M-estimators

5. METHODOLOGY (PART-A)

For the Time Series Analysis, various factors come into play, such as data stationarity, types of models used, and best model selection. Therefore, the Methodology is divided into six parts, namely Time Series Data Decomposition, Box-Cox Transformation, Data Stationarity Tests, Autoregressive Integrated Moving Average (ARIMA) Model, Seasonal Autoregressive Integrated Moving Average (SARIMA) Model and Evaluation Metrics. These sections are elaborated on below:

5.1. Data Decomposition

The Time Series Data (y_t) can be broken up or decomposed into three components, specifically Seasonal (S_t), Trend-Cycle (T_t) and Remainder (R_t) components. There are two types of decomposition i.e., Additive and Multiplicative decomposition, which is explained in detail in Appendix-D under the heading of "Data Decomposition".

The decomposition of time series is utilized to make series stationery by subtracting seasonal and trend cycle components from the main time series. The stationarity of the time series will be explained in the Data Stationarity Tests Section. To implement the Additive or Multiplicative Decomposition decompose Function of the stats Library in the R Programming Language (Sebastien, 2021).

5.2. Box-Cox Transformation

The Box-Cox Transformation comes into use when the time series data shows fluctuations that decrease or increase with the level of the series. The Box-Cox Transformation is made up of two transformations, particularly Power and Logarithmic Transformations, and it is dependent on the λ parameter (Box & Cox, 1964). In Logarithmic Transformation, the original observed values y_t are transformed to w_t values such that $w_t = \log(y_t)$. Furthermore, in Power Transformation, the original observed values y_t are transformed to w_t values such that $w_t = (y_t)^p$ (Rob & George, 2018a). The formula of the Box-Cox Transformation is given in Figure 3.

$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ (y_t^{\lambda} - 1)/\lambda & \text{otherwise.} \end{cases}$$

Figure 3: Formula of the Box-Cox Transform

More detailed information about Box-Cox Transformation is given in Appendix-D under the heading of "Box-Cox Transformation".

5.3. Data Stationarity Tests

The time series data comes in two forms, specifically Stationary and Non-Stationary. A time series whose statistical characteristics are independent of time is called a stationary time series. More specifically, if we suppose a stationary time series{y_t}, then for every s, the distribution (y_t, …, y_{t+s}) is independent or does not dependent on t. Therefore, time series data with seasonality or trend-cycle is considered to be non-stationary time series. In contrast, white noise or random time series is considered a stationary time series (Rob & George, 2021a).

The time series can be made stationary by subtracting seasonal and trend cycle components from the original time series, which is explained in detail in Data Decomposition Section. Furthermore, time series can also be made stationary by performing a Box-Cox Transformation on the original time series to make the variance constant which is explained in detail in Box-Cox Transformation Section.

However, sometimes it is difficult to know whether a time series is stationary or non-stationary just by observing the time series' pattern. To overcome this problem, several Stationary Tests are utilized, which are explained in Appendix-D under the header "Data Stationarity Tests".

5.4. ARIMA Model

To understand the Auto-Regressive Integrated Moving Average (ARIMA) models, we first need to understand the Auto-Regressive (AR) and Moving Average Models (MA). These Models are Explained below:

1) Auto-Regressive (AR) Models:

   In Auto-Regressive Models, we predict the future values of a variable by utilizing a linear combination of past values of the variable. The term Auto-Regressive means that the linear regression is being performed on its own values (Rob & George, 2018b). The formula of the autoregressive model is shown in Figure 4.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

Figure 4: Formula of AR Model where y_{t-p} p times lagged version y_t

   In the above formula, ε_t is white noise, and the y_{t-1} to y_{t-p} are lagged values of y_t. We represent the AR Model by AR(p) (Rob & George, 2018b).

2) Moving Average (MA) Models:

   Unlike AR Model, the Moving Average Model conducts a linear regression of past forecast error values, i.e., the future values are predicted with the help of past error values of the forecast (Rob & George, 2018c). The formula of the moving average model is shown in Figure 5.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

Figure 5: Formula of MA Model where ε_t is the error term

   In the above formula, ε_t is white noise, and the ε_{t-1} to ε_{t-p} are lagged values of ε_t. We represent the MA Model by MA(q) (Rob & George, 2018c).

Another major aspect of ARIMA apart from AR and MA models is differencing term, which is represented as I(d). The differencing is a method in which the time-lagged series of the original time series is subtracted from the original time series to get a stationary series or differenced series (Rob & George, 2021b).

Now, if we concatenate the Auto-Regressive and Moving Average Model with a differencing term, we get a nonseasonal ARIMA Model. The formula of the ARIMA Model is given below in Figure 6.

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

Figure 6: Formula of ARIMA Model with φ and θ as coefficients for AR and MA terms, respectively

Here, $(y_t)'$ represents the differenced series. Furthermore, the terms on the left-hand side of the equation include both lagged error and auto terms. The ARIMA model is represented as ARIMA(p, d, q) (Rob & George, 2021c).

The ARIMA Model is implemented by the Arima and auto.arima functions of forecast library in R Programming Language (Hyndman et. al, 2022).

Now, the selection of p, q and d terms in the ARIMA Model requires the analysis of Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots along with the Box-Ljung test, Akaike's Information Criterion (AIC), Corrected Akaike's Information Criterion (AICc), Schwarz's Bayesian Information Criterion (BIC) and Residual plots. These all are explained in detail in Appendix-D under the heading of "ARIMA Model".

5.5. SARIMA Model

A Seasonal Autoregressive Integrated Moving Average (SARIMA) is an ARIMA Model that accepts seasonal data. Furthermore, SARIMA is created by adding certain seasonal terms in ARIMA Model (Rob & George, 2018e). The representation of the SARIMA Model is shown in Figure 7.

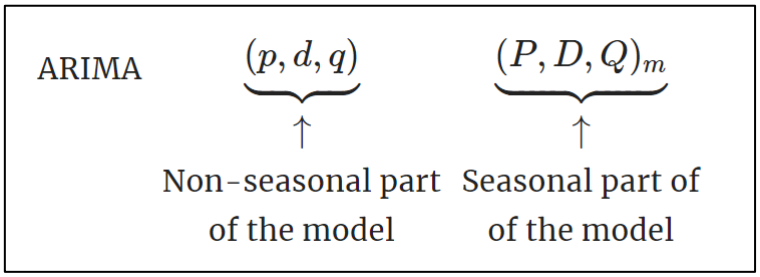

Figure 7: Representation of the SARIMA Model where P, D, and Q are seasonal terms that subtracts actual data values along with error values with their own *m* time-lagged data values and errors

Here, m is the number of observations annually. The terms present in the seasonal part shown in Figure 7 are similar to the Non-Seasonal Terms, but the former terms perform the backshift of the seasonal period (Rob & George, 2018e). For instance, let us consider an ARIMA(1,1,1)(1,1,1)$_4$ model for quarterly data, as shown in Figure 8.

$$(1 - \phi_1 B)\,(1 - \Phi_1 B^4)(1 - B)(1 - B^4) y_t = (1 + \theta_1 B)\,(1 + \Theta_1 B^4)\varepsilon_t.$$

Figure 8: ARIMA(1,1,1)(1,1,1)$_4$ model with B as a Backshift Operator

Here, B is a Backshift Operator, which lags a variable depending on its power; for example, $B^2\ x_t = x_{t-2}$ (PennState, 2022).

The SARIMA is implemented by Sarima and auto.sarima methods in bayesforecast Library in R Programming Language (Matamoros et al.,2021). The ACF, PCF, Box-Ljung Test, AIC, AICc and BIC explained in the ARIMA Model Section are applied in the SARIMA Model in the exact same way as in the ARIMA Model.

An ARIMA(0,0,0)(1,0,0)$_{12}$ will be utilized when the seasonal lags of the ACF plots dampen exponentially, and there is a single measurable spike at the 12th time lag in the plot of PACF (Rob & George, 2018e). An

ARIMA(0,0,0)$(0,0,1)_{12}$ will be utilized when the seasonal lags of the PACF plots dampen exponentially, and there is a measurable spike at the 12^{th} time lag in the plot of ACF (Rob & George, 2018e).

5.6. Evaluation Metrics

In the first part of the study, we will be utilizing RMSE and MAPE as evaluation metrics. The theory of RMSE and MAPE has been explained in the Appendix-D section under the heading of "Evaluation Metrics".

6. EXPLANATORY DATA ANALYSIS (PART-A)

For Explanatory Data Analysis, different ARIMA and SARIMA Models were trained and tested to get the best models for time series analysis (forecasting). All the tested and trained ARIMA and SARIMA models are given in Tables 13 to 36 in the Appendix-A Section of this Report. Moreover, the Forecast of ARIMA along with SARIMA, ACF and PACF plots are also given in the Appendix-A Section in Figures 11 to 46.

The Explanatory Data Analysis Section has two parts, i.e., Results and Discussion. In the Results Section, we have tables displaying the accuracy of the best ARIMA and SARIMA Models. Whereas in the Discussion Section, we will discuss the results we have received. These Sections are explained below in detail:

6.1. Results

The RMSE and MAPE values of the best ARIMA Models for Salinity, Water Level and Water Temperature at Biggara, Albury, Colignan and Murray Bridge, are given below:

Table 1: RMSE and MAPE Score of Best Models for Salinity, Water Level and Water Temperature at Biggara

Location	Parameter	Best ARIMA Model	Model Accuracy (RMSE)	Model Accuracy (MAPE)
Biggara	Salinity	(3,1,1)	16.98	48.08
	Water Level	(3,0,2)	0.09	157.72
	Water Temperature	(1,0,4)	0.92	515.51

Table 2: RMSE and MAPE Score of Best Models for Salinity, Water Level and Water Temperature at Albury

Location	Parameter	Best ARIMA Model	Model Accuracy (RMSE)	Model Accuracy (MAPE)
Albury	Salinity	(1,1,2)	9.25	13.22
	Water Level	(5,0,0)	0.08	84.68
	Water Temperature	(1,0,4)	0.40	147.18

Table 3: RMSE and MAPE Score of Best Models for Salinity, Water Level and Water Temperature at Colignan

Location	Parameter	Best ARIMA Model	Model Accuracy (RMSE)	Model Accuracy (MAPE)
Colignan	Salinity	(1,1,2)	44.81	26.49
	Water Level	(1,0,4)	0.06	38.26
	Water Temperature	(3,0,1)	0.40	218.40

Table 4: RMSE and MAPE Score of Best Models for Salinity, Water Level and Water Temperature at Murray Bridge

Location	Parameter	Best ARIMA Model	Model Accuracy (RMSE)	Model Accuracy (MAPE)
Murray Bridge	Salinity	(2,1,0)	73.79	28.88
	Water Level	(1,0,2)	0.05	254.96
	Water Temperature	(3,0,2)	0.34	109.17

The RMSE and MAPE values of the best SARIMA Models for Salinity, Water Level and Water Temperature at Biggara, Albury, Colignan and Murray Bridge, are given below:

Table 5: RMSE and MAPE Score of Best Models for Salinity, Water Level and Water Temperature at Biggara

Location	Parameter	Best SARIMA Model	Model Accuracy (RMSE)	Model Accuracy (MAPE)
Biggara	Salinity	(5,0,1)(0,1,0)[365]	20.49	53.66
	Water Level	(2,1,2)(0,1,0)[365]	0.14	8.17
	Water Temperature	(1,0,2)(0,1,0)[365]	1.16	5.94

Table 6: RMSE and MAPE Score of Best Models for Salinity, Water Level and Water Temperature at Albury

Location	Parameter	Best SARIMA Model	Model Accuracy (RMSE)	Model Accuracy (MAPE)
Albury	Salinity	(3,0,0)(0,1,0)[365]	11.1	15.19
	Water Level	(2,1,2)(0,1,0)[365]	0.11	4.53
	Water Temperature	(2,0,2)(0,1,0)[365]	0.56	2.94

Table 7: RMSE and MAPE Score of Best Models for Salinity, Water Level and Water Temperature at Colignan

Location	Parameter	Best SARIMA Model	Model Accuracy (RMSE)	Model Accuracy (MAPE)
Colignan	Salinity	(5,0,0)(0,1,0)[365]	43.72	24.77
	Water Level	(3,1,5)(0,1,0)[365]	0.07	1.54
	Water Temperature	(1,0,5)(0,1,0)[365]	0.53	1.92

Table 8: RMSE and MAPE Score of Best Models for Salinity, Water Level and Water Temperature at Murray Bridge

Location	Parameter	Best SARIMA Model	Model Accuracy (RMSE)	Model Accuracy (MAPE)
Murray Bridge	Salinity	(0,0,5)(0,1,0)[365]	46.18	15.76
	Water Level	(1,1,3)(0,1,0)[365]	0.06	7.00
	Water Temperature	(2,0,2)(0,1,0)[365]	0.45	1.17

6.2. Discussion

There were three research questions asked in the first part of this study, which we will be answering by analysing Tables 1 to 8 above. Each of the Research Questions is answered and discussed below in detail:

1) Research Question 1 (Salinity Forecast):

   The forecast provided by the ARIMA model shows that the trend for 178 days is constant for all four locations, which can be seen in Figures 11, 20, 29 and 38 in Appendix-A. However, when we look at the forecast given by the SARIMA model in Figures 14 and 23 for 178 days, we discover that the trend first goes up and then down for Murray Bridge and Colignan, which indicates that there is some seasonality. In contrast, the forecast of the SARIMA model for Albury and Biggara in Figures 32 and 41 had a constant trend with some fluctuations for 178 days.

   The RMSE and MAPE values of the SARIMA model were less than those of the ARIMA model, which indicates that the SARIMA Model has outperformed the ARIMA Model for Albury and Biggara. Whereas for Murray Bridge and Colignan ARIMA Model is better than the SARIMA Model because ARIMA has lower RMSE and MAPE values than SARIMA.

2) Research Question 2 (Water Level Forecast):

   The forecast generated by the ARIMA model shows that the trend for 178 days is downward for Albury, Biggara, and Murray Bridge, which can be seen in Figures 30, 39 and 12, respectively. Whereas for Colignan, ARIMA's forecast shows an upward trend in Figure 21. The projections given by the SARIMA model in Figures 33 and 15 show an overall upward trend for Albury and Murray Bridge, respectively, whereas it is downward and constant with fluctuations for Colignan and Biggara, respectively, as shown in Figures 24 and 42.

   If we compare the difference in RMSE values between ARIMA and SARIMA models with the difference in MAPE values between ARIMA and SARIMA models, we observe that difference in MAPE values is far greater than that of RMSE values. Therefore, for all four locations, the best Model is SARIMA.

3) Research Question 3 (Water Temperature Forecast):

   The forecast provided by the ARIMA model shows that the trend for 178 days is downward for all four locations, which can be seen in Figures 13, 22, 31 and 40. In comparison, the forecast generated by the SARIMA model showed that the trend is overall in an upward direction for all four locations, which can be seen in Figures 16, 25, 34 and 43.

   When we examine, in contrast, the difference in RMSE values between ARIMA and SARIMA models with the difference in MAPE values between ARIMA and SARIMA models, we notice that

the difference in MAPE values is far greater than that of RMSE values. Therefore, for all four locations, the best Model is SARIMA.

Lastly, we got to know that SARIMA is the best model for Water Level and Water Temperature for all four locations. Whereas for Salinity, ARIMA was best for Albury and Biggara, and SARIMA was best for Colignan and Murray Bridge.

Now, because the MAPE and RMSE values given by both ARIMA (at Albury and Biggara) and SARIMA (at Colignan and Murray Bridge) for Salinity were very large when compared to those of the SARIMA models of Water Level and Water Temperature, we would need to find another more accurate model for Salinity.

However, it should be noted that for getting faster Salinity forecasts, ARIMA can be a preferred choice for Albury and Biggara. Similarly, for Water Level at Albury and Colignan, ARIMA can be utilized for faster projections because their RMSE and MAPE values are close to that of SARIMA. Whereas Water Level ARIMA for Biggara and Murray Bridge have MAPE values about 20 folds greater than those of SARIMA. For Water Temperature at all four locations, SARIMA should be preferred.

According to Abdullah et al. (2021) and Amirkhalili et al., (2020) Support Vector Machine (SVM) model performed better than SARIMA and ARIMA models, respectively. Therefore, in the second part of the research, our focus will be on the SVM method.

7. CONCLUSION (PART-A)

The first part of the research project paper presented a comparison between two statistical models i.e., ARIMA and SARIMA Models for Salinity, Water Level and Water Temperature time series data provided by MDBA Authority for four locations, specifically Biggara, Albury, Colignan and Murray Bridge. It was observed that projections generated by SARIMA displayed seasonal behaviour for Salinity at Colignan and Murray Bridge. However, the Salinity forecast by ARIMA was constant throughout Albury and Biggara. For Water Level, the projections had a declining trend for Albury, Biggara, and Murray Bridge. In contrast, the Water Level forecast by SARIMA at Colignan had an increasing trend. Finally, Water Temperature predictions given by the SARIMA model had an upward pattern.

Overall, in terms of forecast performance, SARIMA was the best model for Water Level and Water Temperature when observing the difference values of RMSE and MAPE values of ARIMA and SARIMA. In comparison, the forecast efficiency of ARIMA was better for Salinity at Albury and Biggara. Whereas, for Salinity at Colignan and Murray Bridge, SARIMA gave a better forecast efficiency. Finding the best model was necessary to create a better Forecasting Model in Future to prevent disastrous situations in the MDB Basin, which has an enormous area and is a significant area for a lot of endangered species for their survival.

8. BACKGROUND (PART-B)

Lin et al. (2006) have presented the Support Vector Machine (SVM) as a model for forecasting hydrological variables. The investigators claim that the SVM model is immune to over-fitting and locally optimal solutions because the SVM model utilizes the structural risk minimization principle instead of the empirical risk minimization principle (Lin et al., 2006). Lin et al. (2006) tested the SVM model with the aid of long-term monthly streamflow data provided by the Manwan Hydropower Complex. The results of the study indicated that the SVM Model performed better than Auto-Regressive and Moving Average (ARMA) and Artificial Neural Network (ANN) models in forecasting long-term streamflow data (Lin et al., 2006).

In the second part of the research, we would be only forecasting Salinity because Water Level and Water Temperature were forecasted more accurately than Salinity by their respective SARIMA Models. This is because the RMSE and MAPE values of Water Level along with Water Temperature were very small when compared to that of Salinity. Thus, to improve the accuracy of the Salinity forecast, we would be utilizing the SVM model, which is a machine-learning method.

More importantly, we would be using the best SARIMA models of Water Level and Water Temperature, which are discussed above in the first part of the research, to forecast time-series data of both variables from July 2022 to July 2023 to help the SVM model in forecasting Salinity. The Water Temperature, Water Level and Time and Date variables would be utilized as predictor variables along with Time and Date variable, for the response variable, Salinity. Another reason for selecting Water Level and Water Temperature as predictor variables is because they are very uncorrelated to Salinity, as shown in Tables 38, 40, 42 and 44.

9. RESEARCH QUESTIONS (PART-B)

The Research Questions (for the second part of the research) for the Time Series Analysis of the Murray Darling Basin (MDB) River System are:

1) Which model, ARIMA or SARIMA or SVM, is better for performing a Time-Series Analysis (Forecasting) for Salinity?

10. METHODOLOGY (PART-B)

For the Machine Learning (ML) Time Series Analysis, various factors come into play, such as hyperparameter optimization, types of kernels used, and best model selection. Therefore, the Methodology is divided into three parts, namely Hyperparameter Optimization, Support Vector Machine (SVM), Radial basis function (RBF) and Evaluation Metrics. These sections are elaborated on below:

10.1. Support Vector Machine

Support Vector Machine (SVM) is an ML algorithm that locates a hyperdimensional plane that segregates different classes (Agrawal, 2021). An instance of the model is shown in Figure 9, where there are two classes designated by blue and red colours. In addition, the black line separating these two classes is called a hyperplane which can be seen clearly in Figure 10. Furthermore, SVM locates the hyperplane by maximizing the distance between two dotted lines shown in Figure 9, which is known as the margin (Agrawal, 2021).

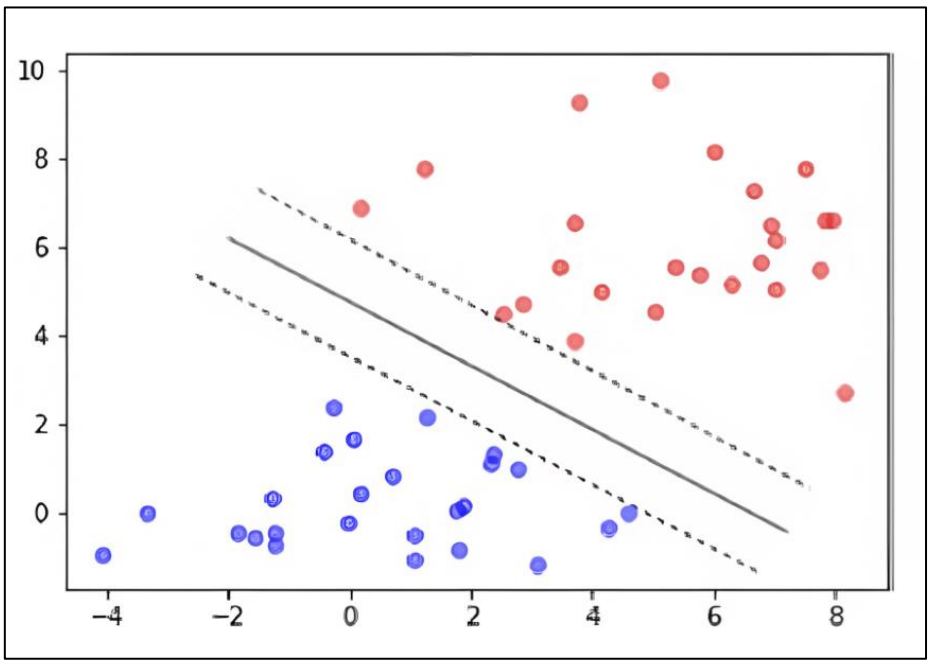

Figure 9: Two distinct classes segregated by a hyperplane in the second dimension

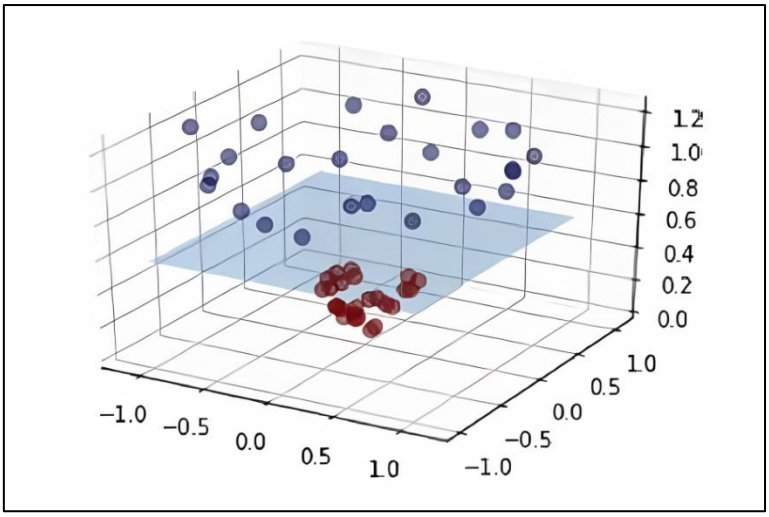

Figure 10: Data projected into the third dimension where two distinct classes are segregated by a hyperplane

The above explanation of SVM is for classification. However, in this study, we will be utilizing SVM for regression. The mathematical explanation of SVM for regression is provided in detail in Appendix-E under the heading "Support Vector Machine".

10.2. Radial Basis Function

The explanation of the kernel function and the reasons for selecting the Radial Basis Function are provided in Appendix-E under the heading "Radial Basis Function".

10.3. Hyperparameter Optimization

In every machine learning (ML) algorithm, there are two types of variables specifically:

1) Parameters:
   These are the variables that are optimized by the machine learning model in accordance with the provided dataset (Agrawal, 2021).

2) Hyperparameters:

   These are the higher-level variables that can be set manually by the user before the machine-learning model is trained (Agrawal, 2021).

The RBF kernel of the SVM model contains three hyperparameters, namely cost or regularization constant (C), margin (ε) and sigma (σ). Furthermore, it should be noted that $2\sigma^2 = 1/\gamma$, where γ is gamma.

In this research, we would be only tuning hyperparameters C and σ due to simplicity. Now, to find the optimal value of C and σ, we will conduct a parameter search on the training data by performing v-fold cross-validation on a grid that contains different values of C and σ.

According to Hsu et al. (2003), in v-fold cross-validation, we divide our training set into v equally sized subsets. Each of the subsets is consecutively tested on the RBF SVM model, which was trained using the remaining $(v - 1)$ subsets (Hsu et al., 2003).

In our research, we perform 5-fold cross-validation on different values of C and σ present in the grid on 5 equal subsets of training data. This method is referred to as Grid search (Hsu et al., 2003). The top 5 values of C and σ are selected based on the lowest values of RMSE, MAPE and Standard Deviation (SD).

The v-fold cross-validation is performed by the vfold_cv function of the rsample library and grid search by the tune_grid function of the tune library of R Programming Language (Silge, 2022; Kuhn, 2022b). The plots of grid search using 5-fold cross-validation are given in Appendix-C from Figure 55 to Figure 58. Furthermore, the Top-5 values of C and σ are given in Appendix-B in Tables 37, 39, 41 and 43.

10.4. Evaluation Metrics

In the second part of the study, we will be utilizing RMSE and MAPE as evaluation metrics. The theory of RMSE and MAPE has been explained in the Appendix-D section under the heading “Evaluation Metrics”.

11. EXPLANATORY DATA ANALYSIS (PART-B)

For Explanatory Data Analysis, different RBF SVM Models were trained and tested to get the best models for time series analysis (forecasting) of Salinity. All the tested and trained RBF SVM models are given in Tables 37, 39, 41 and 43 in the Appendix-B Section of this Report. Moreover, the Salinity Forecast of ARIMA along with SARIMA, ACF and PACF plots are also given in the Appendix-A Section in Figures 11 to 46. Furthermore, the Salinity Forecast of RBF SVM is given in Figures 47 to 54 in Appendix-B.

The Explanatory Data Analysis Section has two parts, i.e., Results and Discussion. In the Results Section, we have tables displaying the accuracy of the best RBF SVM models. Whereas in the Discussion Section, we will discuss the results we have received. These Sections are explained below in detail:

11.1. Results

The RMSE and MAPE values of the best SVM RBF Models for Salinity as response variable (with Water Level and Water Temperature as predictor variables) at Biggara, Albury, Colignan and Murray Bridge is given below:

Table 9: RMSE and MAPE Score of RBF SVM in comparison with ARIMA and SARIMA for Salinity at Albury

Location	Best Parameters	Model	MAPE	RMSE
Albury	(1,1,2) = (p,d,q)	ARIMA	13.22	9.25
	(3,0,0)(0,1,0)[365] = (p,d,q)(P,D,Q)[m]	SARIMA	15.19	11.1
	(1,1) = (σ,C)	RBF SVM	11.51	7.86

Table 10: RMSE and MAPE Score of RBF SVM in comparison with ARIMA and SARIMA for Salinity at Biggara

Location	Best Parameters	Model	MAPE	RMSE
Biggara	(3,1,1) = (p,d,q)	ARIMA	48.08	16.98
	(5,0,1)(0,1,0)[365] = (p,d,q)(P,D,Q)[m]	SARIMA	53.66	20.49
	(1,3.17) = (σ,C)	RBF SVM	44.41	16.8

Table 11: RMSE and MAPE Score of RBF SVM in comparison with ARIMA and SARIMA for Salinity at Colignan

Location	Best Parameters	Model	MAPE	RMSE
Colignan	(1,1,2) = (p,d,q)	ARIMA	26.49	44.81
	(5,0,0)(0,1,0)[365] = (p,d,q)(P,D,Q)[m]	SARIMA	24.77	43.72
	(1,0.31) = (σ,C)	RBF SVM	17.99	35.88

Table 12: RMSE and MAPE Score of RBF SVM in comparison with ARIMA and SARIMA for Salinity at Murray Bridge

Location	Best Parameters	Model	MAPE	RMSE
Murray Bridge	(2,1,0) = (p,d,q)	ARIMA	28.88	73.79
	(0,0,5)(0,1,0)[365] = (p,d,q)(P,D,Q)[m]	SARIMA	15.76	46.18
	(0.75,0.1,0.01) = (σ,C, ε)	RBF SVM	15.53	46.07

11.2. Discussion

There was one research question asked in the second part of this study, which we will be answering by analysing Tables 9 to 12 above. This Research Question is answered and discussed below in detail:

The response variable for RBF SVM is Salinity, with Water Temperature, and Water Level as predictor variables because SVM had drastic performance change when predictor variables were increased in number. However, ARIMA and SARIMA both did not display any efficiency change when covariates were introduced.

The Salinity forecast provided by the ARIMA model shows that the trend for 365 days is constant for all four locations, which can be seen in Figures 11, 20, 29 and 38. However, when we look at the Salinity forecast given by the SARIMA model in Figures 14 and 23 for 365 days, we discover that the trend first goes up and then down for Murray Bridge and Colignan, which indicates that there is some seasonality. In contrast, the forecast of the SARIMA model for Albury and Biggara in Figures 32 and 41 had a constant trend with some fluctuations for 365 days. At last, the Salinity projections generated by RBF SVM at Albury and Biggara in Figures 52 and 54 show a gradually increasing trend with fluctuations. However, the forecast at Colignan in Figure 50 displays a constant trend with some variations. In addition, the trend of Salinity predictions at Murray Bridge in Figure 48 displayed first an upward trend and then a declining trend which informs us of some seasonal behaviour.

Overall, in terms of forecast efficiency, the RBF SVM had the lowest values of RMSE and MAPE when compared to both ARIMA and SARIMA for all four locations. Although for Murray Bridge, there was only a slight improvement in forecast performance by SVM, probably because of more seasonality in Salinity.

However, it should be noted that for getting faster Salinity forecasts, ARIMA can be a preferred choice for Albury and Biggara. Also, for Murray Bridge, SARIMA and RBF SVM can be interchangeably used because their RMSE and MAPE values are quite close.

12. CONCLUSION (PART-B)

The second part of the research project paper presented a comparison between two statistical models i.e., ARIMA and SARIMA Models and one machine learning model, namely RBF SVM, for Salinity as a response variable and Water Level along with Water Temperature as predictor variables. The above time series data was provided by MDBA Authority for four locations, specifically Biggara, Albury, Colignan and Murray Bridge. It was observed that projections generated by SARIMA displayed seasonal behaviour for Salinity at Colignan and Murray Bridge. However, the Salinity forecast by ARIMA was constant throughout Albury and Biggara. Finally, the Salinity forecast by RBF SVM presented that at Albury and Biggara, the trend of forecast gradually increased. In comparison, it remained constant at Colignan with few fluctuations, and at Murray bridge, it had a seasonal behaviour.

Overall, in terms of forecast performance, RBF SVM was the best model for Salinity because its RMSE and MAPE values were least when compared with ARIMA and SARIMA. Finding the best model was necessary to create a better Forecasting Model in Future to prevent disastrous situations in the MDB Basin, which has an enormous area and is a significant area for a lot of endangered species for their survival.

13. FUTURE WORK

In future work, following new models can be explored to handle variability and seasonality in data. The Gated Recurrent Unit (GRU) and Long-Term Short Memory (LSTM) Deep learning models can be utilized as they outperform ARIMA and SARIMA (ArunKumar et al., 2022). According to Mateus et al. (2021), LSTM and GRU have displayed superior efficiency while forecasting sequential data. Where sequential data represents data points which are dependent on other data points in the dataset. More significantly, LSTM is able to draw out the patterns from sequential data and retain these patterns in the internal state of variables, thus making it good at time series forecasting. Furthermore, every LSTM cell being utilized has the ability to store valuable information for a longer time. This enables LSTM to perform efficiently in predicting, processing and classifying convoluted dynamic sequences (Mateus et al., 2021). Therefore, in future work to improve the forecast performance of the Salinity, one can take the help of GRU or LSTM.

REFERENCES

Abdullah, A., Ruchjana, B., Jaya, M., & Soemartini, S. (2021). Comparison of SARIMA and SVM model for rainfall forecasting in Bogor city, Indonesia. *Journal of Physics: Conference Series*, *1722*, 012061. https://doi.org/10.1088/1742-6596/1722/1/012061

Agrawal, T. (2021). Introduction to hyperparameters. In *Hyperparameter Optimization in Machine Learning* (pp. 1–30). Apress. https://doi.org/10.1007/978-1-4842-6579-6$_1$

Amirkhalili, Y., Aghsami, A., & Jolai, F. (2020). *Comparison of time series ARIMA model and support vector regression*. *13*, 12. https://doi.org/10.21742/ijhit.2020.13.1.02

ArunKumar, K. E., Kalaga, D. V., Kumar, C. M. S., Kawaji, M., & Brenza, T. M. (2022). Comparative analysis of Gated Recurrent Units (GRU), long Short-Term memory (LSTM) cells, autoregressive Integrated moving average (ARIMA), seasonal autoregressive Integrated moving average (SARIMA) for forecasting COVID-19 trends. *Alexandria Engineering Journal*, *61*(10), 7585-7603.

Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the american statistical association*, *76*(374), 296-311.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211-243.

Cao, L. J., & Tay, F. E. H. (2003). Support vector machine with adaptive parameters in financial time series forecasting. IEEE Transactions on Neural Networks, 14(6), 1506–1518. https://doi.org/10.1109/tnn.2003.820556

Chen, P., Niu, A., Liu, D., Jiang, W., & Ma, B. (2018). Time series forecasting of temperatures using SARIMA: An example from Nanjing. *IOP Publishing*, *394*, 052024.

Department for Environment and Water. (2022). *Taking action on Salinity*. Department for Environment and Water. Retrieved August 19, 2022, from https://www.environment.sa.gov.au/topics/river-murray/improving-river-health/water-quality-and-salinity/taking-action-on-salinity

Fox, J., & Weisberg, S. (2013, October 8). *Robust regression - School of Statistics*. Robust regression. Retrieved August 15, 2022, from http://users.stat.umn.edu/~sandy/courses/8053/handouts/robust.pdf

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). *A Practical Guide to Support Vector Classification*. https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Hyndman, R. J. (2017). *Forecasting: Principles and practice (UWA)*. 2.4 Non-seasonal ARIMA models. Retrieved August 16, 2022, from https://robjhyndman.com/uwa2017/2-4-NonseasonalARIMA.pdf

Hyndman, R. J., & Athanasopoulos, G. (2018a). *Forecasting: Principles and practice (2nd ed)*. 6.1 Time series components. Retrieved August 15, 2022, from https://otexts.com/fpp2/components.html

Hyndman, R. J., & Athanasopoulos, G. (2018b). *Forecasting: Principles and practice (2nd ed)*. 8.3 Autoregressive models. Retrieved August 16, 2022, from https://otexts.com/fpp2/AR.html

Hyndman, R. J., & Athanasopoulos, G. (2018c). *Forecasting: Principles and practice (2nd ed)*. 8.4 Moving average models. Retrieved August 16, 2022, from https://otexts.com/fpp2/MA.html

Hyndman, R. J., & Athanasopoulos, G. (2018d). *Forecasting: Principles and practice (2nd ed)*. 2.8 Autocorrelation. Retrieved August 15, 2022, from https://otexts.com/fpp2/autocorrelation.html

Hyndman, R. J., & Athanasopoulos, G. (2018e). *Forecasting: Principles and practice (2nd ed)*. 8.9 Seasonal ARIMA models. Retrieved August 15, 2022, from https://otexts.com/fpp2/seasonal-arima.html

Hyndman, R. J., & Athanasopoulos, G. (2018f). *Forecasting: Principles and practice (2nd ed)*. 3.4 Evaluating forecast accuracy. Retrieved November 17, 2022, from https://otexts.com/fpp2/accuracy.html

Hyndman, R. J., & Athanasopoulos, G. (2021a). *Forecasting: Principles and practice (3rd ed)*. 9.1 Stationarity and differencing. Retrieved August 16, 2022, from https://otexts.com/fpp3/stationarity.html#fn15

Hyndman, R. J., & Athanasopoulos, G. (2021b). *Forecasting: Principles and practice (3rd ed)*. 9.1 Stationarity and differencing. Retrieved August 16, 2022, from https://otexts.com/fpp3/stationarity.html

Hyndman, R. J., & Athanasopoulos, G. (2021c). *Forecasting: Principles and practice (3rd ed)*. 9.5 Non-seasonal ARIMA models. Retrieved August 16, 2022, from https://otexts.com/fpp3/non-seasonal-arima.html

Hyndman, R. J., & Athanasopoulos, G. (2021d). *Forecasting: Principles and practice (3rd ed)*. 7.5 Selecting predictors. Retrieved August 16, 2022, from https://otexts.com/fpp3/selecting-predictors.html

Hyndman, R. J., & Athanasopoulos, G. (2021e). *Forecasting: Principles and practice (3rd ed)*. 5.4 Residual diagnostics. Retrieved August 16, 2022, from https://otexts.com/fpp3/diagnostics.html#diagnostics

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., Kuroptev, K., O'Hara-Wild, M., & Petropoulos, F. (2022, July 25). *Forecasting functions for time series and linear models [R package forecast*

version 8.17.0]. The Comprehensive R Archive Network. Retrieved August 16, 2022, from https://cran.r-project.org/web/packages/forecast/index.html

Keerthi, S. S., & Lin, C.-J. (2003). Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. *Neural Computation*, *15*(7), 1667–1689. https://doi.org/10.1162/089976603321891855

Kokoszka, P., & Young, G. (2016). *KPSS test for Functional Time Series*. Colorado State University. Retrieved August 16, 2022, from https://www.stat.colostate.edu/~piotr/kpss.pdf

Kuhn, M. (2022a). *Parsnip: A common API to modeling and analysis functions*. Package 'parsnip.' Retrieved November 21, 2022, from https://cran.r-project.org/web//packages//parsnip/parsnip.pdf

Kuhn, M. (2022b). *Tune: Tidy tuning tools*. Package 'tune.' Retrieved November 21, 2022, from https://cran.r-project.org/web//packages//tune/tune.pdf

Lin, H. T., & Lin, C. J. (2003). A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. *Neural Comput*, *3*(1-32), 16.

LIN, J.-Y., CHENG, C.-T., & CHAU, K.-W. (2006). Using support vector machines for long-term discharge prediction. *Hydrological Sciences Journal*, *51*(4), 599–612. https://doi.org/10.1623/hysj.51.4.599

Loo, M. van der. (2022, June 16). *simputation Package*. Getting started with simputation. Retrieved August 15, 2022, from https://cran.r-project.org/web/packages/simputation/vignettes/intro.html

Matamoros, A. A., Torres, C. C., Dala, A., Hyndman, R., & O'Hara-Wild, M. (2021, June 17). *Bayesian time series modeling with Stan [R package bayesforecast version 1.0.1]*. The Comprehensive R Archive Network. Retrieved August 16, 2022, from https://cran.microsoft.com/snapshot/2022-05-26/web/packages/bayesforecast/index.html

Mateus, B. C., Mendes, M., Farinha, J. T., Assis, R., & Cardoso, A. M. (2021). Comparing LSTM and GRU Models to Predict the Condition of a Pulp Paper Press. *Energies*, *14*(21), 6958.

Murray Darling Basin Authority. (2022a). *About the data: River Murray Data*. Live River Data. Retrieved August 15, 2022, from https://riverdata.mdba.gov.au/about-data

Murray Darling Basin Authority. (2022b). *Salinity*. Australian Government - Murray-Darling Basin Authority. Retrieved August 15, 2022, from https://www.mdba.gov.au/issues-murray-darling-basin/salinity

Murray Darling Basin Authority. (2022c). *Acid sulfate soils*. Australian Government - Murray-Darling Basin Authority. Retrieved August 15, 2022, from https://www.mdba.gov.au/issues-murray-darling-basin/acid-sulfate-soils

Murray-Darling Basin Authority. (2022d). *Drought*. Australian Government - Murray-Darling Basin Authority. Retrieved August 15, 2022, from https://www.mdba.gov.au/issues-murray-darling-basin/drought

Murray-Darling Basin Authority. (2022e). *The murray–darling basin and why it's important*. Australian Government - Murray-Darling Basin Authority. Retrieved August 16, 2022, from https://www.mdba.gov.au/importance-murray-darling-basin

Murray-Darling Basin Authority. (2022f). *Environmental importance*. Australian Government - Murray-Darling Basin Authority. Retrieved August 16, 2022, from https://www.mdba.gov.au/importance-murray-darling-basin/environment

Musarat, M. A., Alaloul, W. S., Rabbani, M. B. A., Ali, M., Altaf, M., Fediuk, R., Vatin, N., Klyuev, S., Bukhari, H., Sadiq, A., Rafiq, W., & Farooq, W. (2021). Kabul River Flow Prediction Using Automated ARIMA Forecasting: A Machine Learning Approach. *Sustainability*, *13*(19), 10720. https://doi.org/10.3390/su131910720

PennState. (2022). *2.3 notational conventions: Stat 510*. PennState: Statistics Online Courses. Retrieved August 16, 2022, from https://online.stat.psu.edu/stat510/lesson/2/2.3

Qiu, D. (2015). Package aTSA. CRAN. Retrieved August 16, 2022, from https://cran.r-project.org/web/packages/aTSA/aTSA.pdf

Silge, J. (2022). *General Resampling Infrastructure [R package rsample version 1.1.0]*. Package 'rsample.' Retrieved November 21, 2022, from https://cran.r-project.org/web/packages/rsample/index.html

Statsmodels. (2022). Stationarity and detrending (ADF/KPSS). Retrieved August 16, 2022, from https://www.statsmodels.org/dev/examples/notebooks/generated/stationarity_detrending_adf_kpss.html

Vapnik, V. (1998). The Support Vector Method of Function Estimation. *Nonlinear Modeling*, *XVII*(978-1-4615-5703-6), 55–85. https://doi.org/10.1007/978-1-4615-5703-6_3

Wouters, S. (2021). Package 'DecomposeR.' Retrieved August 15, 2022, from https://cran.r-project.org/web/packages/DecomposeR/DecomposeR.pdf

Zivot, E. (2014). *Unit root tests*. University of Washington. Retrieved August 16, 2022, from https://faculty.washington.edu/ezivot/econ584/notes/unitroot.pdf

APPENDIX - A

The plot of Murray Bridge Location:

Figure 11: Salinity Levels 365 days Forecast given by the best ARIMA(2,1,0) Model

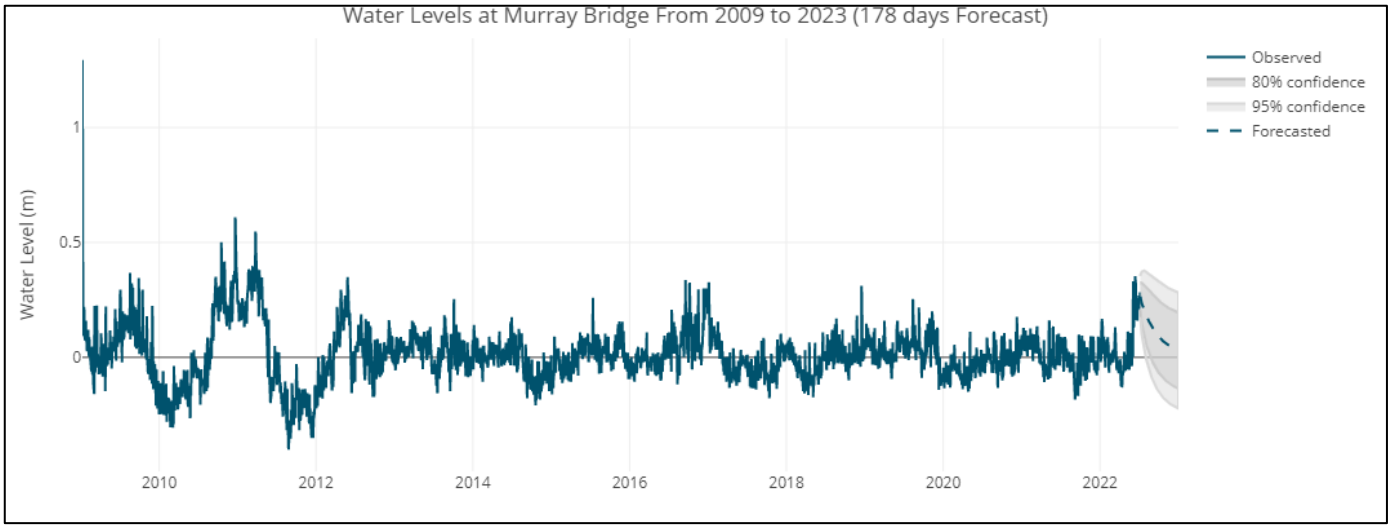

Figure 12: Water Levels 178 days Forecast given by the best ARIMA(1,0,2) Model

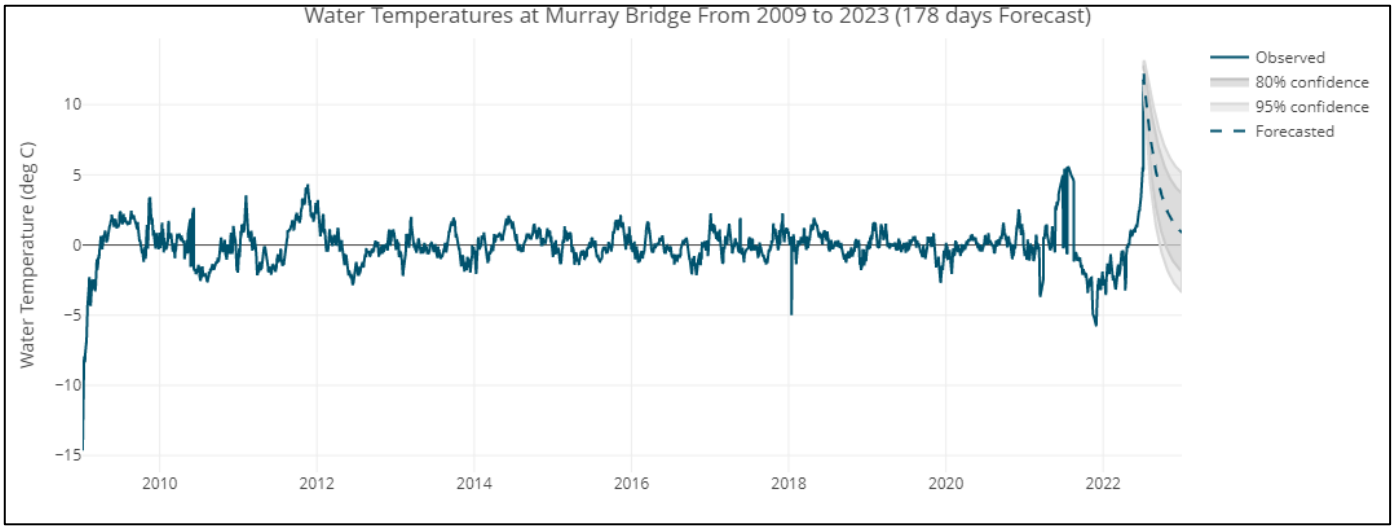

Figure 13: Water Temperatures 178 days Forecast given by the best ARIMA(3,0,2) Model

Figure 14: Salinity Levels 365 days Forecast given by the best ARIMA(0,0,5)(0,1,0)[365] (SARIMA) Model

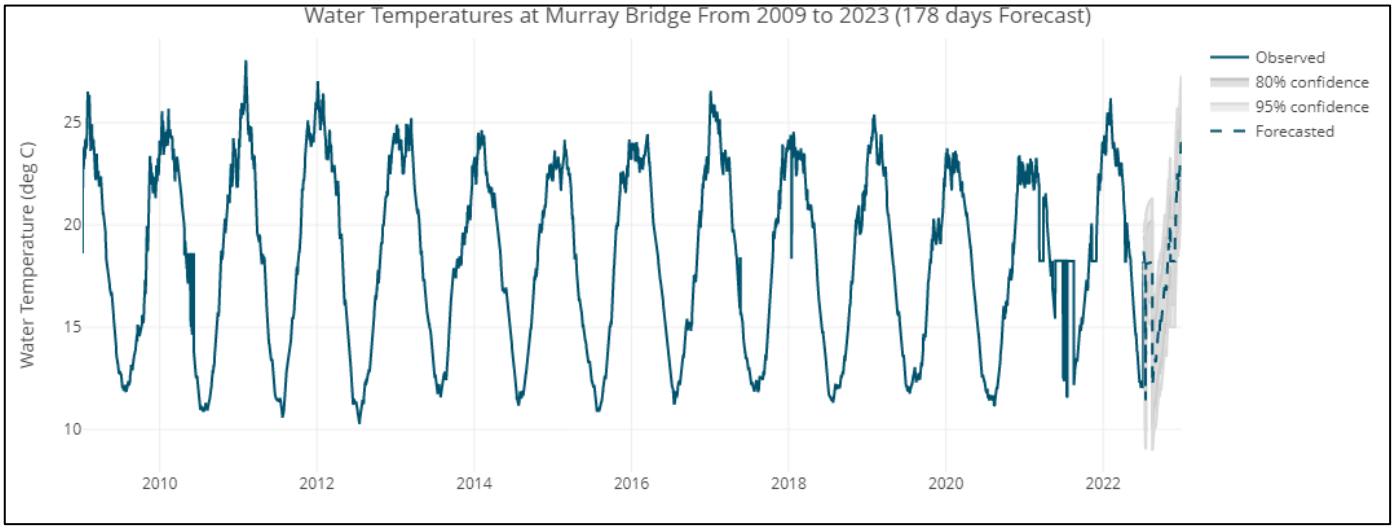

Figure 15: Water Temperatures 178 days Forecast given by the best ARIMA(2,0,2)(0,1,0)[365] (SARIMA) Model

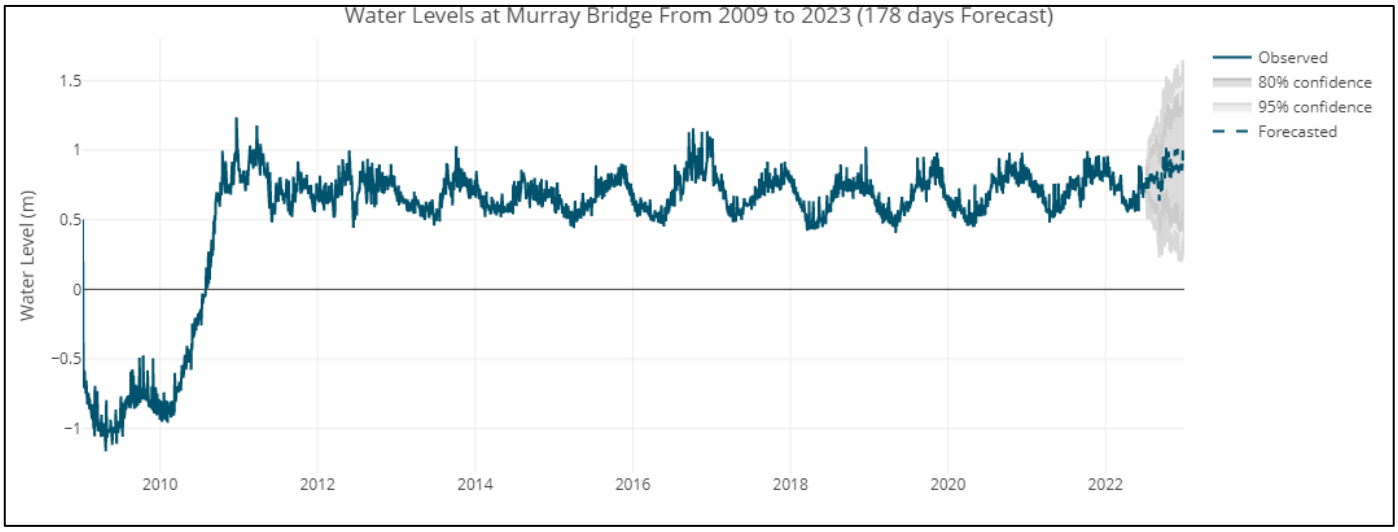

Figure 16: Water Levels 178 days Forecast given by the best ARIMA(1,1,3)(0,1,0)[365] (SARIMA) Model

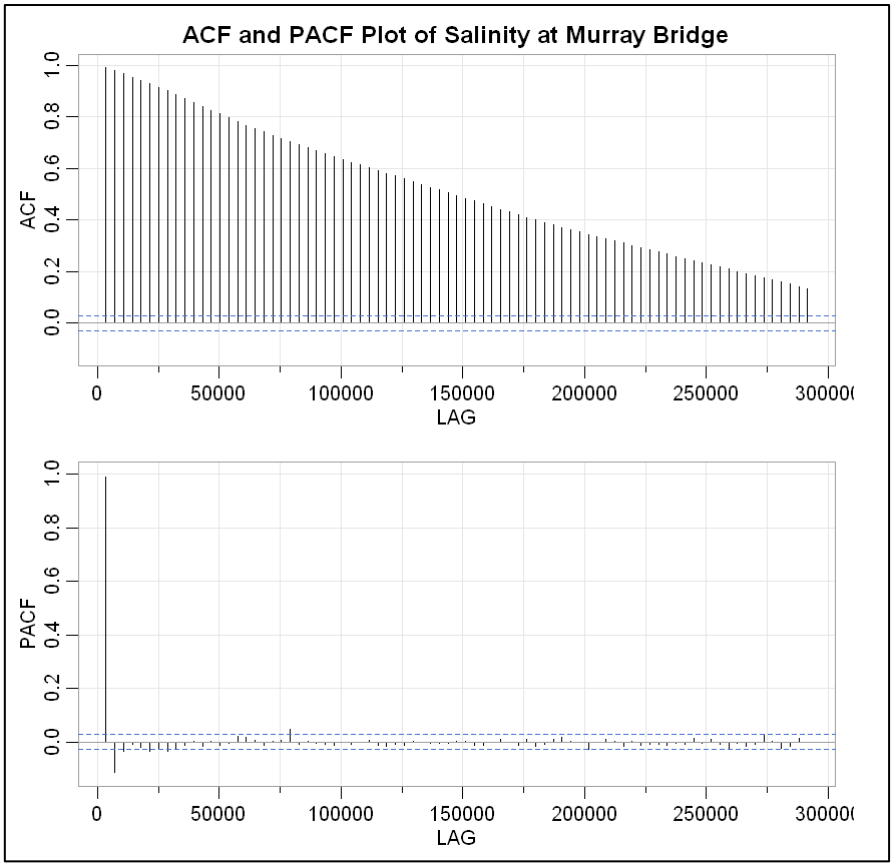

Figure 17: ACF and PACF of Salinity at Murray Bridge. Here, ACF display's a seasonal behaviour

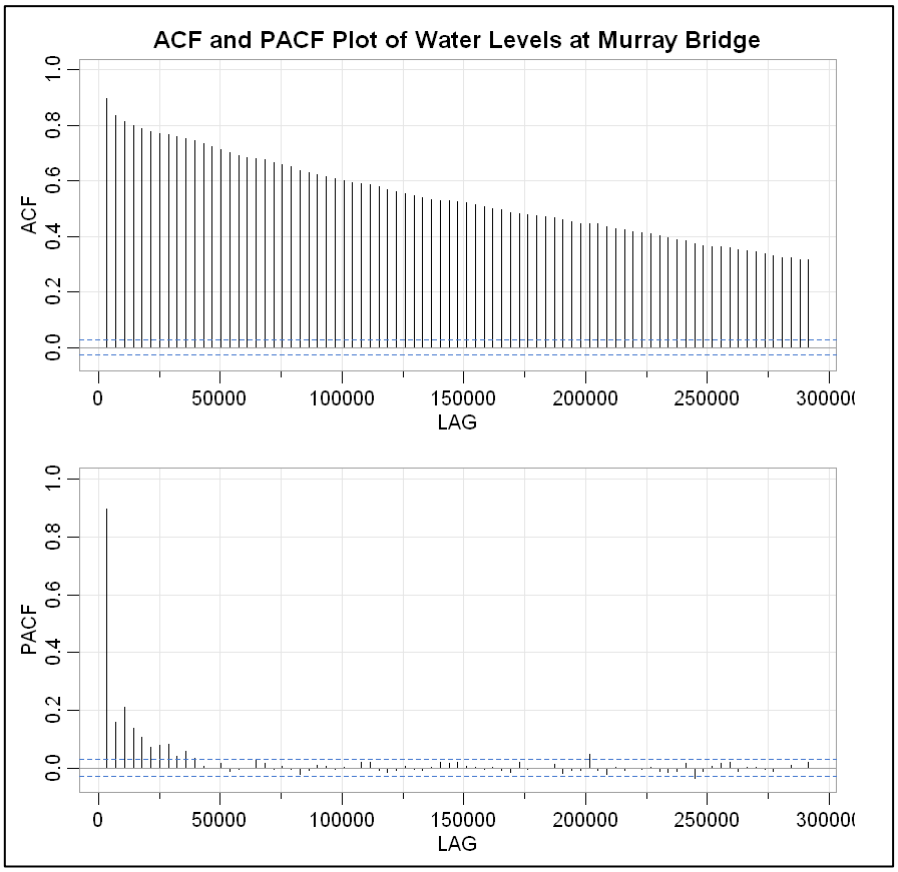

Figure 18: ACF and PACF of Water Levels at Murray Bridge. Here, ACF and PACF display a seasonal behaviour

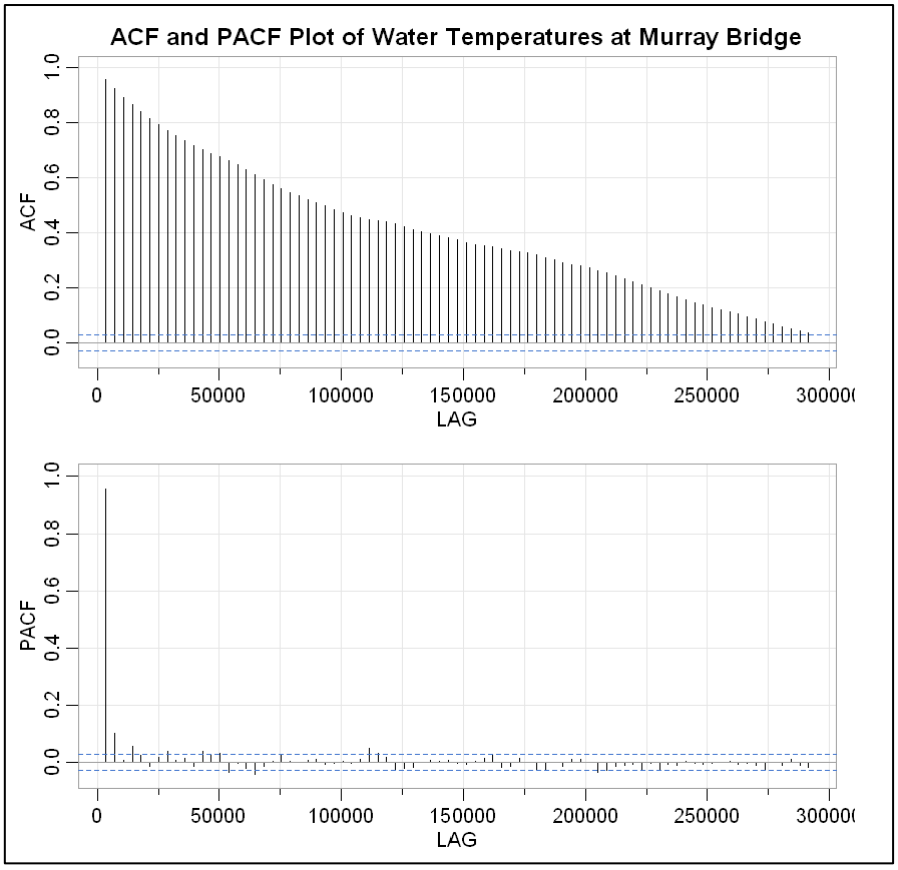

Figure 19: ACF and PACF of Water Temperatures at Murray Bridge. Here, ACF display's a seasonal behaviour

The table data of Murray Bridge Location:

Table 13: Different ARIMA Models Trained for Salinity

ARIMA Models	AIC	AICc	BIC
ARIMA(2,1,0)	35752.55	35752.55	35772.09
ARIMA(0,1,0)	36842.87	36842.87	36849.39
ARIMA(1,1,0)	35803.61	35803.61	35816.64
ARIMA(3,1,0)	35754.07	35754.07	35780.12

Table 14: Different (Top 5) ARIMA Models Trained for Water Level

ARIMA Models	AIC	AICc	BIC
ARIMA(1,0,1)	-16044.42	-16044.41	-16024.91
ARIMA(1,0,2)	-16405.48	-16405.47	-16379.47
ARIMA(2,0,0)	-16003.85	-16003.84	-15984.34
ARIMA(4,0,0)	-16298.26	-16298.25	-16265.75

Table 15: Different (Top 5) ARIMA Models Trained for Water Temperature

ARIMA Models	AIC	AICc	BIC
ARIMA(1,0,1)	3522.903	3522.908	3542.41
ARIMA(1,0,2)	3524.601	3524.609	3550.61
ARIMA(3,0,1)	3515.23	3515.242	3547.741
ARIMA(3,0,2)	3489.25	3489.27	3528.26

Table 16: Different (Top 5) SARIMA Models Trained for Salinity

SARIMA Models	AICc
ARIMA(0,0,2)(0,1,0)[365]	45742
ARIMA(0,0,3)(0,1,0)[365]	42661.05
ARIMA(0,0,4)(0,1,0)[365]	40676.47
ARIMA(0,0,5)(0,1,0)[365]	39259.17

Table 17: Different (Top 5) SARIMA Models Trained for Water Level

SARIMA Models	AICc
ARIMA(1,1,3)(0,1,0)[365]	-10504.28
ARIMA(2,1,2)(0,1,0)[365]	-10504.16
ARIMA(2,1,4)(0,1,0)[365]	-10501.41
ARIMA(3,1,2)(0,1,0)[365]	-10502.08

Table 18: Different (Top 5) SARIMA Models Trained for Water Temperature

SARIMA Models	AICc
ARIMA(1,0,1)(0,1,0)[365]	5606.28
ARIMA(1,0,2)(0,1,0)[365]	5604.333
ARIMA(2,0,1)(0,1,0)[365]	5549.63
ARIMA(2,0,2)(0,1,0)[365]	5549.041

The plot of Colignan Location:

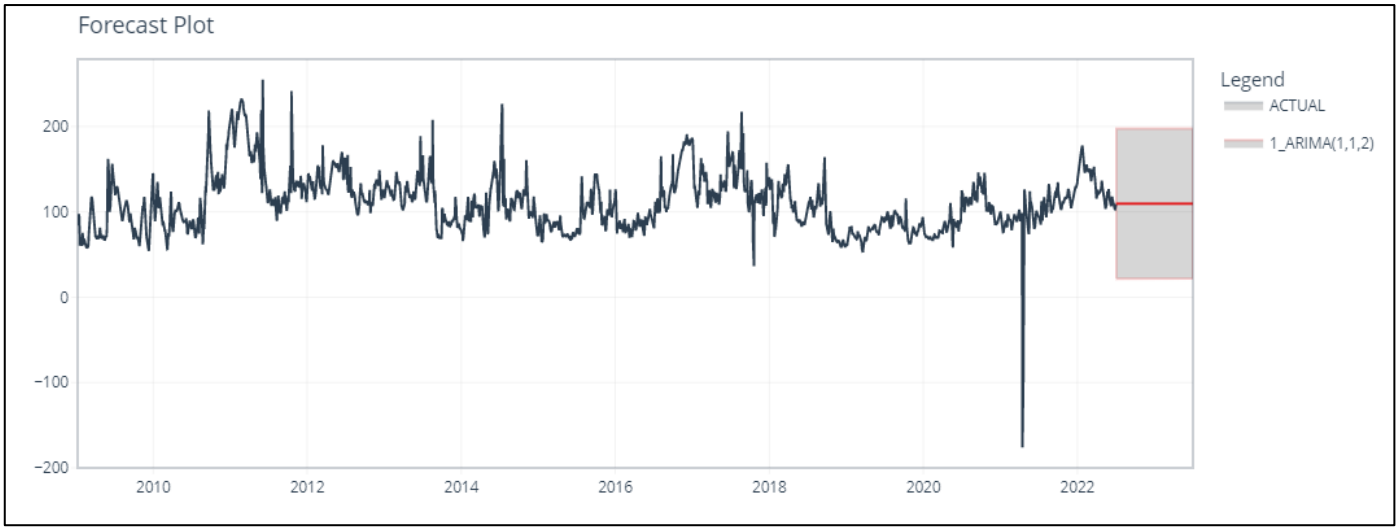

Figure 20: Salinity Levels 365 days Forecast given by the best ARIMA(1,1,2) Model

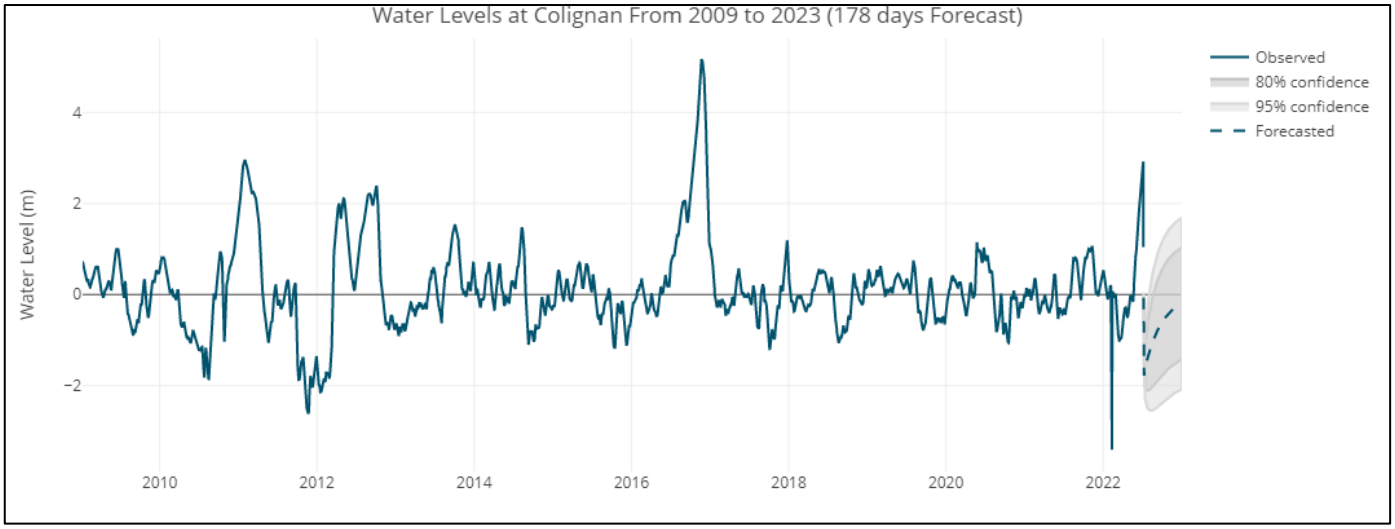

Figure 21: Water Levels 178 days Forecast given by the best ARIMA(1,0,4) Model

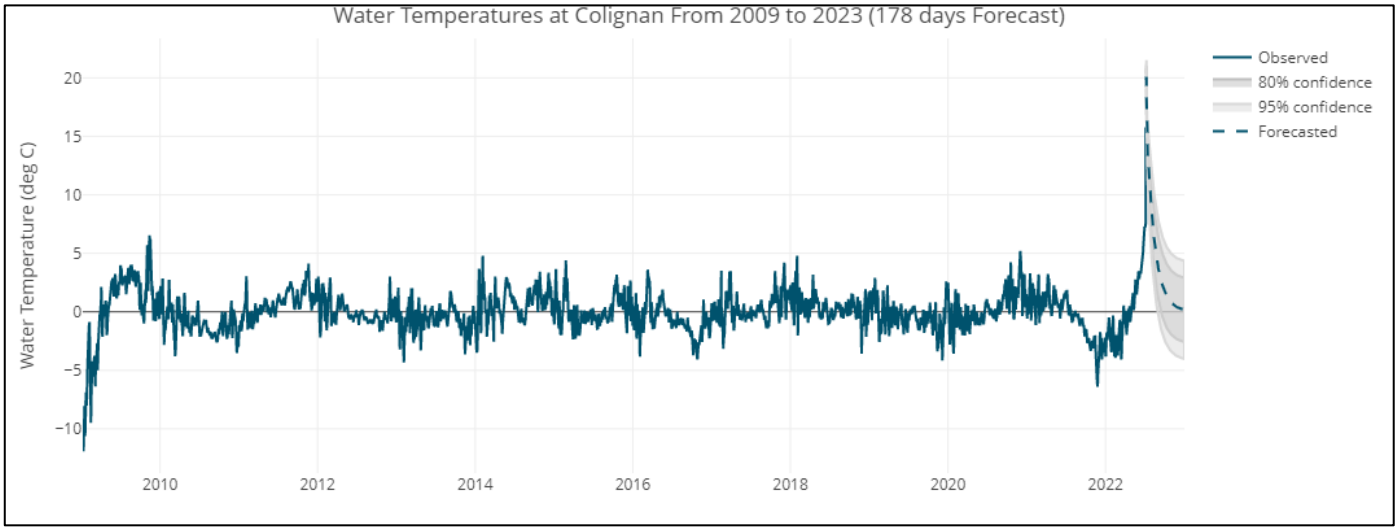

Figure 22: Water Temperatures 178 days Forecast given by the best ARIMA(3,0,1) Model

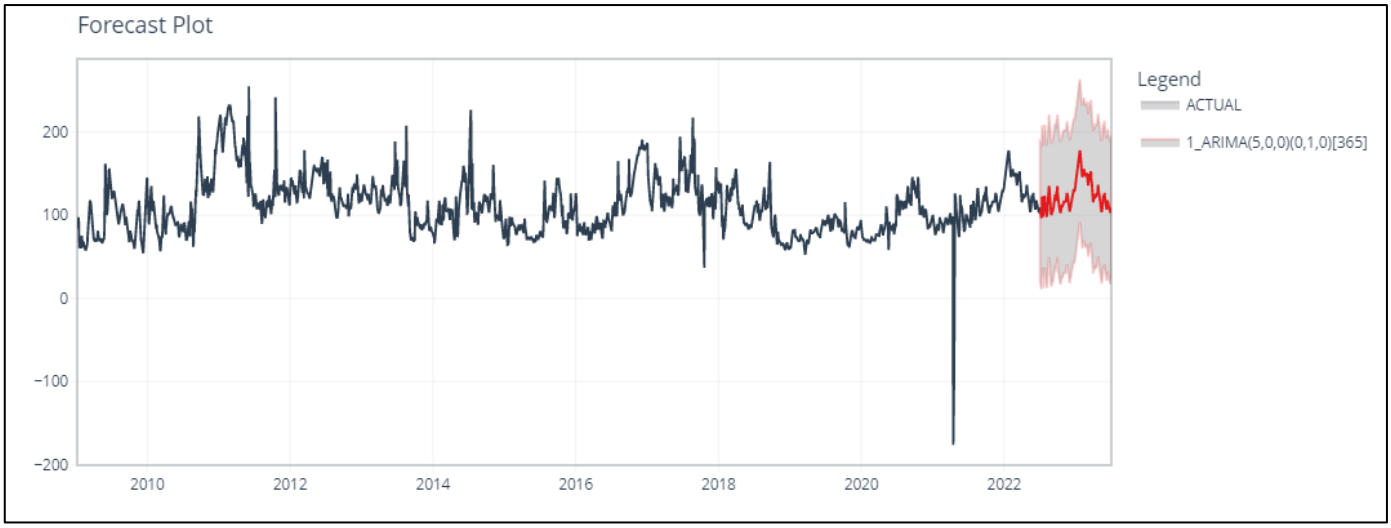

Figure 23: Salinity Levels 365 days Forecast given by the best ARIMA(5,0,0)(0,1,0)[365] (SARIMA) Model

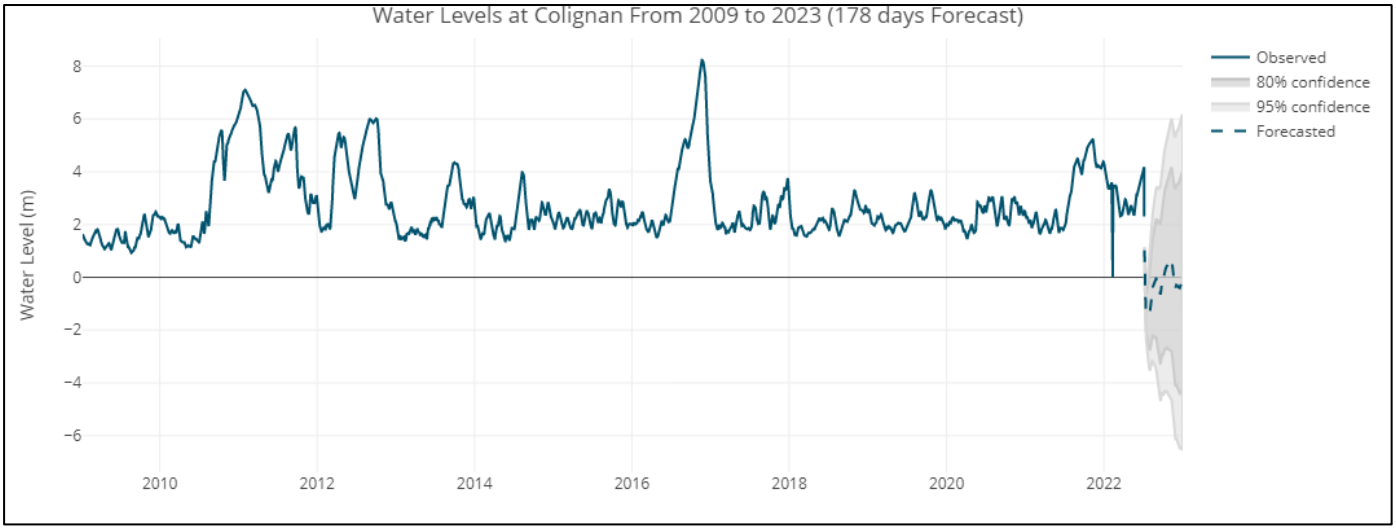

Figure 24: Water Levels 178 days Forecast given by the best ARIMA(3,1,5)(0,1,0)[365] (SARIMA) Model

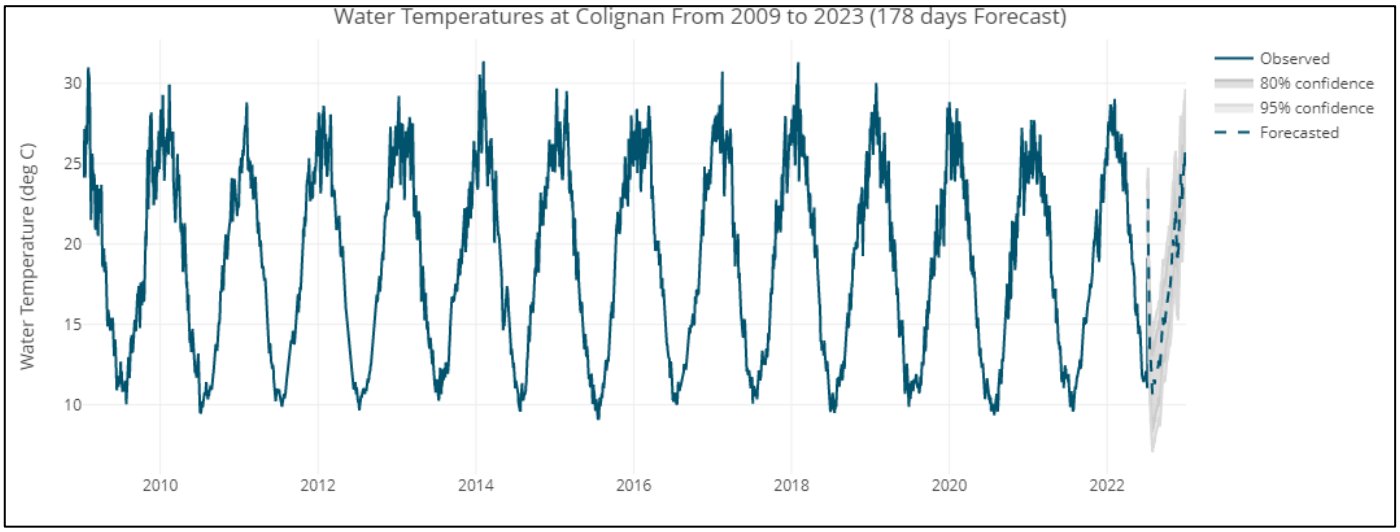

Figure 25: Water Temperatures 178 days Forecast given by the best ARIMA(1,0,5)(0,1,0)[365] (SARIMA) Model

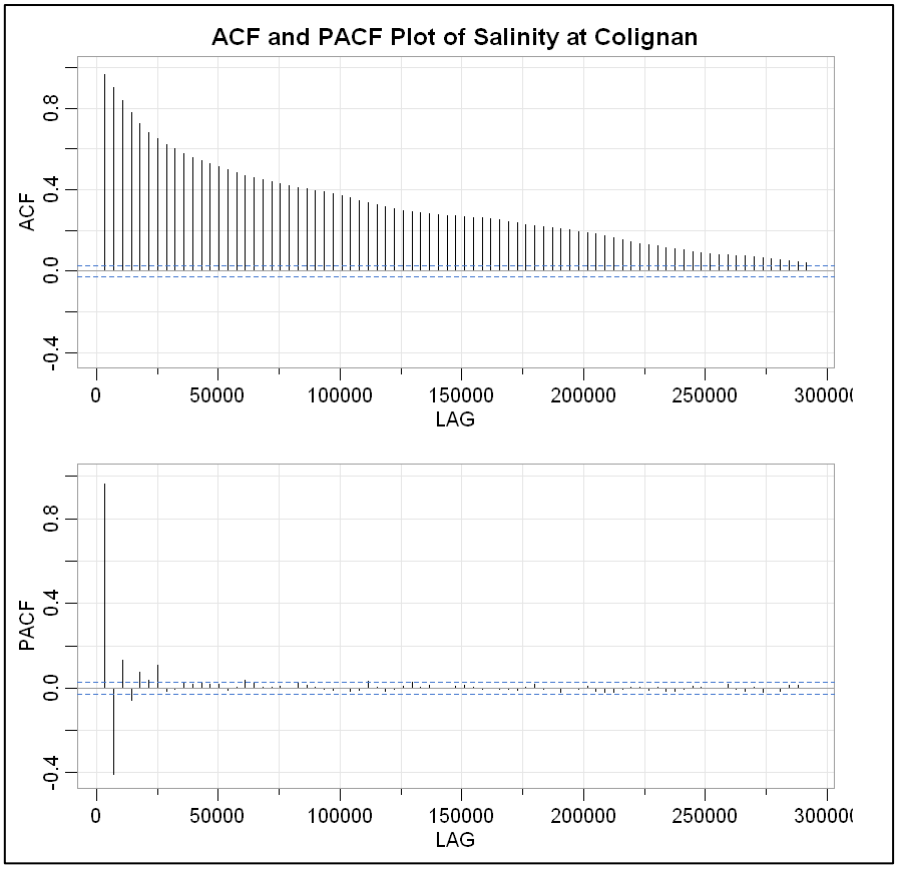

Figure 26: ACF and PACF of Salinity at Colignan. Here, ACF display's a seasonal behaviour

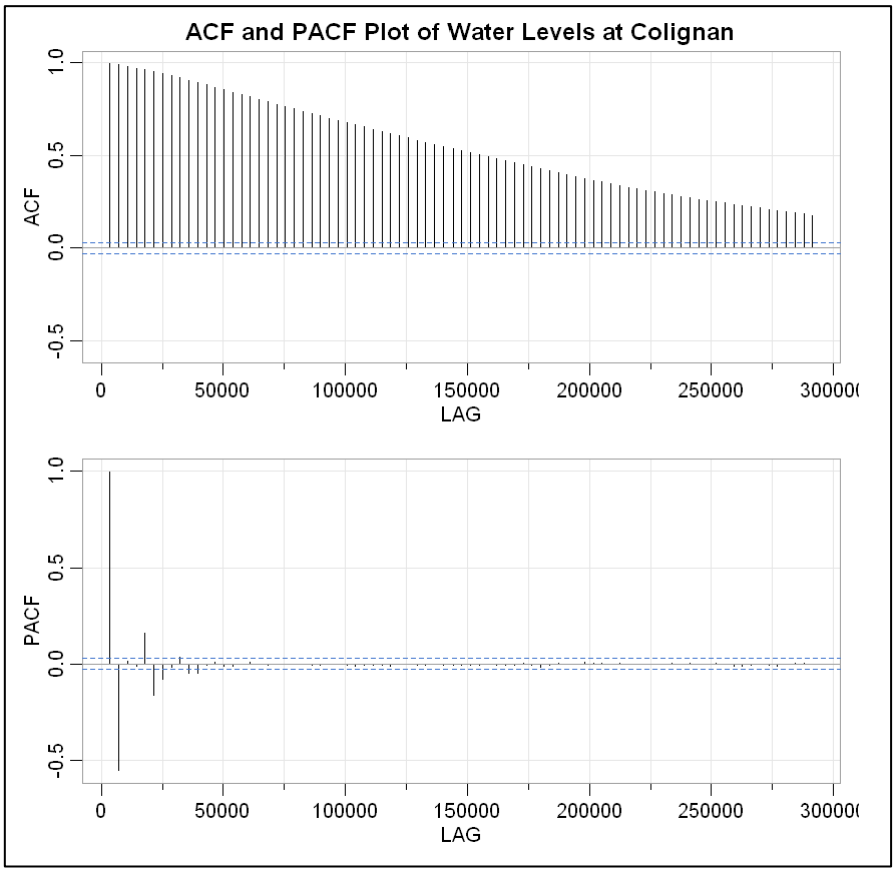

Figure 27: ACF and PACF of Water Levels at Colignan. Here, ACF display's a seasonal behaviour

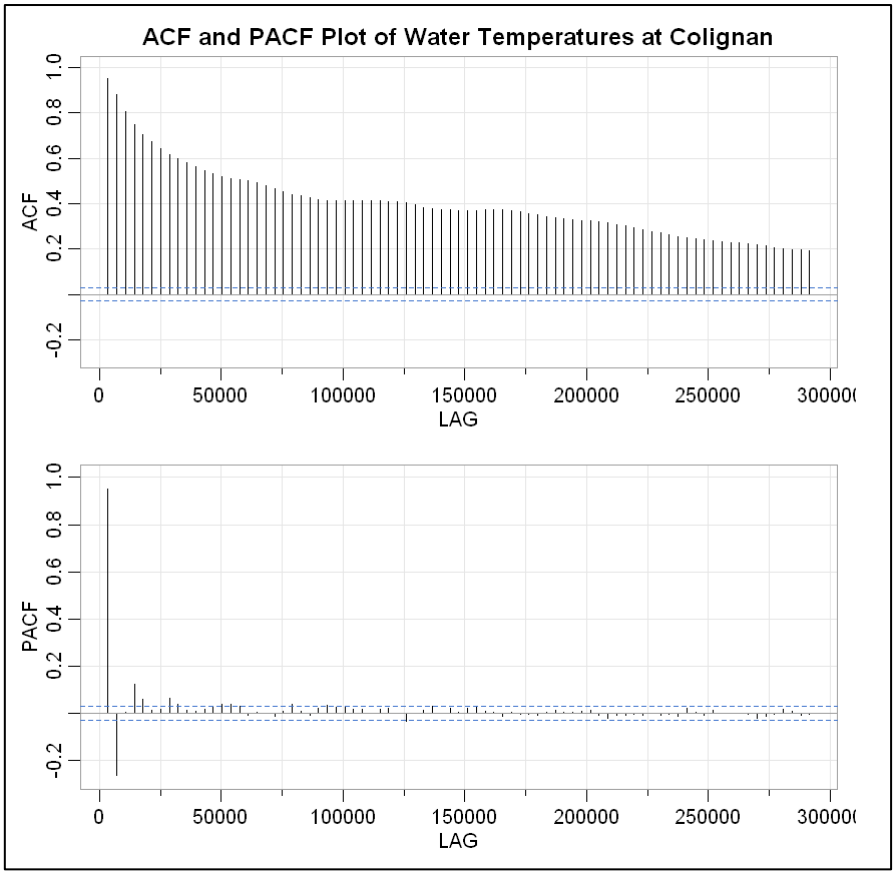

Figure 28: ACF and PACF of Water Temperatures at Colignan. Here, ACF display's a seasonal behaviour

The table data of Colignan Location:

Table 19: Different (Top 5) ARIMA Models Trained for Salinity

ARIMA Models	AIC	AICc	BIC
ARIMA(1,1,2)	26910.93	26910.94	26936.52
ARIMA(0,1,0)	27906.6	27906.6	27913
ARIMA(1,1,0)	27267.16	27267.16	27279.95
ARIMA(0,1,1)	26994.07	26994.07	27006.86

Table 20: Different (Top 5) ARIMA Models Trained for Water Level

ARIMA Models	AIC	AICc	BIC
ARIMA(1,0,3)	-13497.75	-13497.74	-13465.24
ARIMA(1,0,4)	-13505.02	-13505	-13466
ARIMA(2,0,3)	-13501.59	-13501.57	-13462.57
ARIMA(3,0,2)	-13148.45	-13148.44	-13109.44

Table 21: Different (Top 5) ARIMA Models Trained for Water Temperature

ARIMA Models	AIC	AICc	BIC
ARIMA(1,0,3)	5120.625	5120.637	5161.655
ARIMA(1,0,4)	5079.161	5079.178	5118.175
ARIMA(3,0,1)	5026.96	5026.97	5059.47
ARIMA(4,0,0)	5033.769	5033.781	5066.281

Table 22: Different (Top 5) SARIMA Models Trained for Salinity

SARIMA Models	AICc
ARIMA(5,0,0)(0,1,0)[365]	29738.49
ARIMA(4,0,0)(0,1,0)[365]	29773.66
ARIMA(3,0,0)(0,1,0)[365]	29784.28
ARIMA(2,0,0)(0,1,0)[365]	29891.71

Table 23: Different (Top 5) SARIMA Models Trained for Water Level

SARIMA Models	AICc
ARIMA(1,1,0)(0,1,0)[365]	-9323.48
ARIMA(0,1,1)(0,1,0)[365]	-8724.894
ARIMA(3,1,5)(0,1,0)[365]	-9659.549
ARIMA(5,1,4)(0,1,0)[365]	-9656.908

Table 24: Different (Top 5) SARIMA Models Trained for Water Temperature

SARIMA Models	AICc
ARIMA(1,0,4)(0,1,0)[365]	7157.303
ARIMA(1,0,5)(0,1,0)[365]	7123.257
ARIMA(2,0,4)(0,1,0)[365]	7167.581
ARIMA(2,0,5)(0,1,0)[365]	7125.017

The plot of Albury Location:

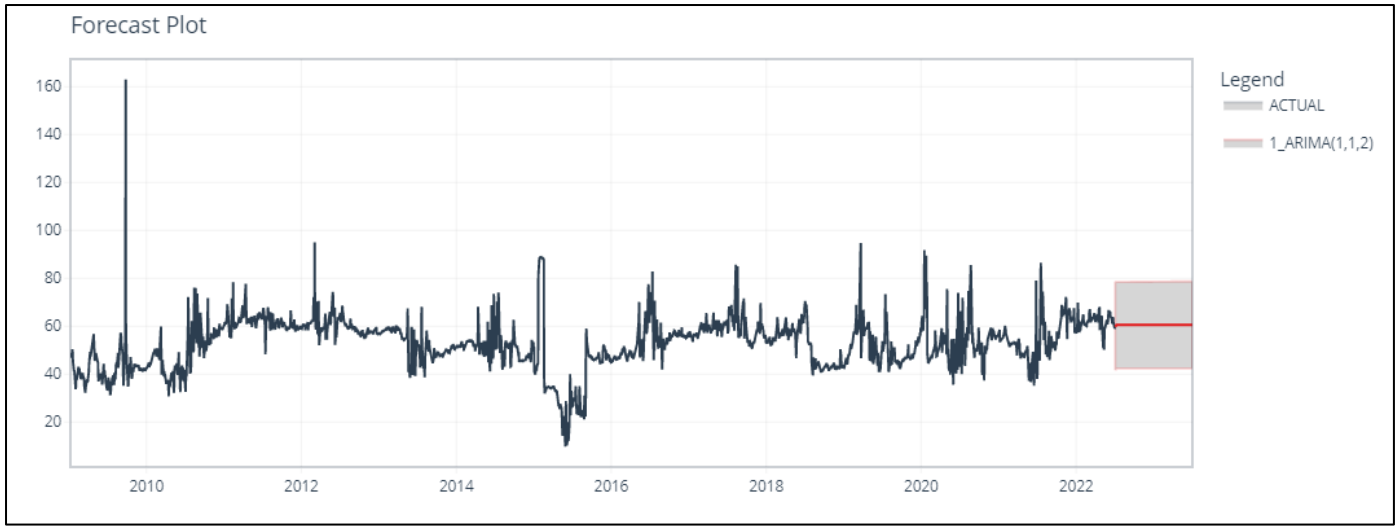

Figure 29: Salinity Levels 365 days Forecast given by the best ARIMA(1,1,2) Model

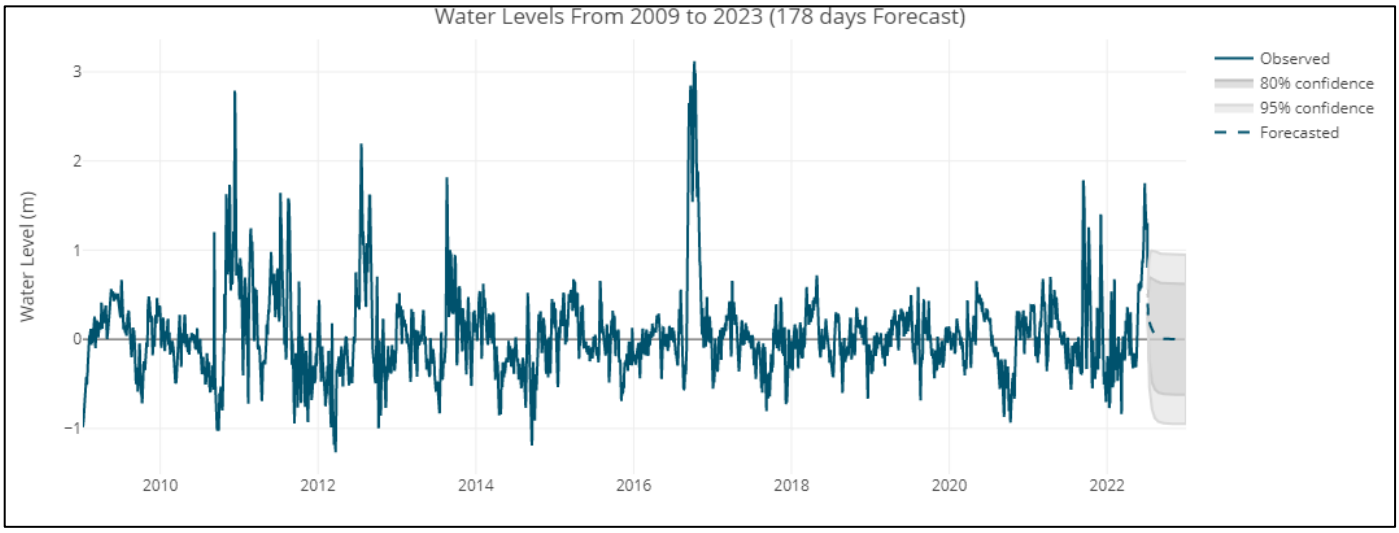

Figure 30: Water Levels 178 days Forecast given by the best ARIMA(5,0,0) Model

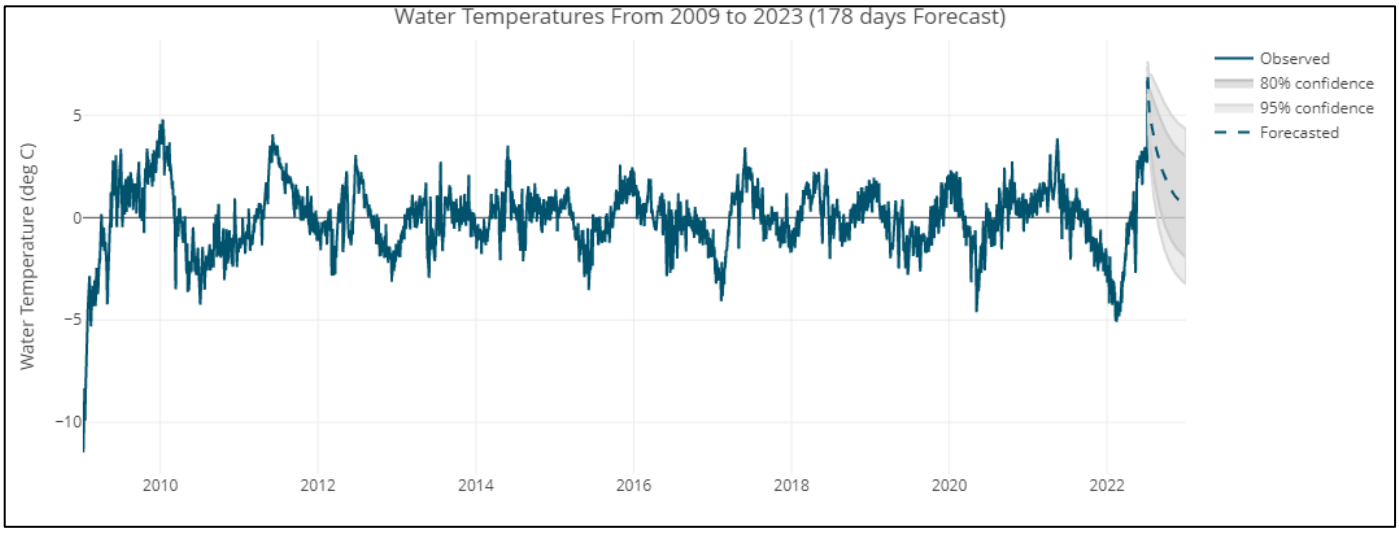

Figure 31: Water Temperatures 178 days Forecast given by the best ARIMA(1,0,4) Model

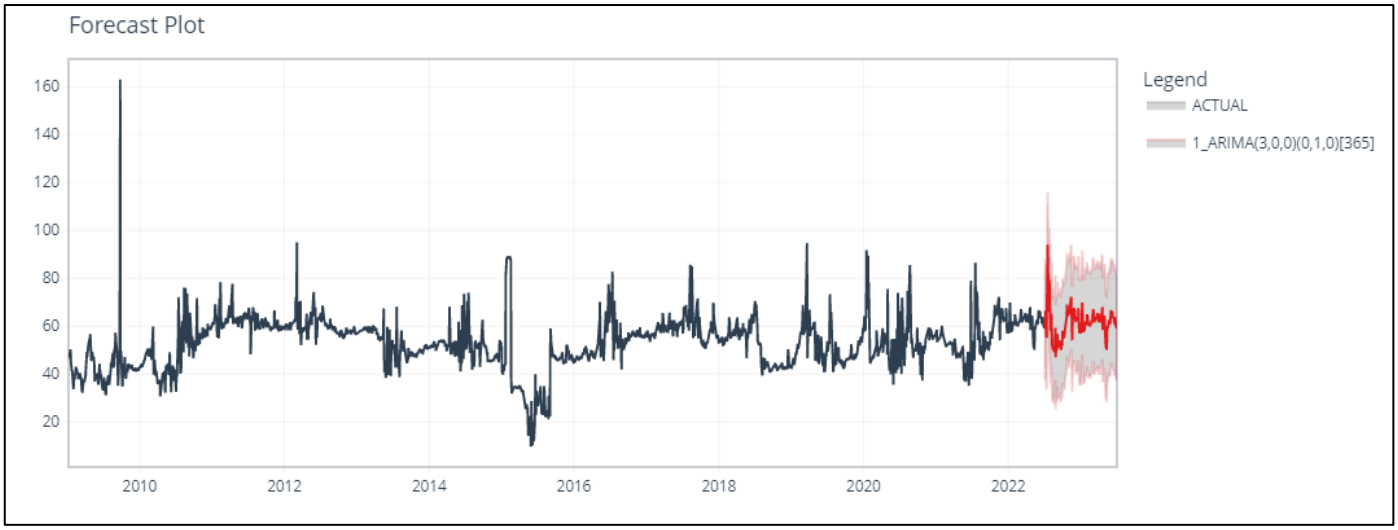

Figure 32: Salinity Levels 365 days Forecast given by the best ARIMA(3,0,0)(0,1,0)[365] (SARIMA) Model

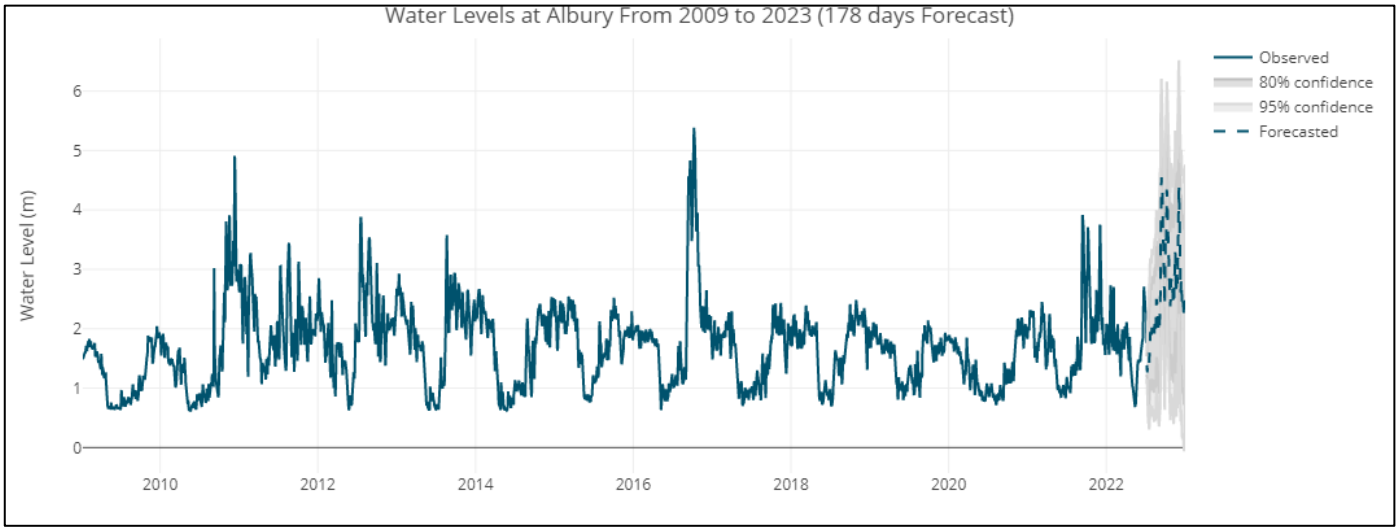

Figure 33: Water Levels 178 days Forecast given by the best ARIMA(2,1,2)(0,1,0)[365] (SARIMA) Model

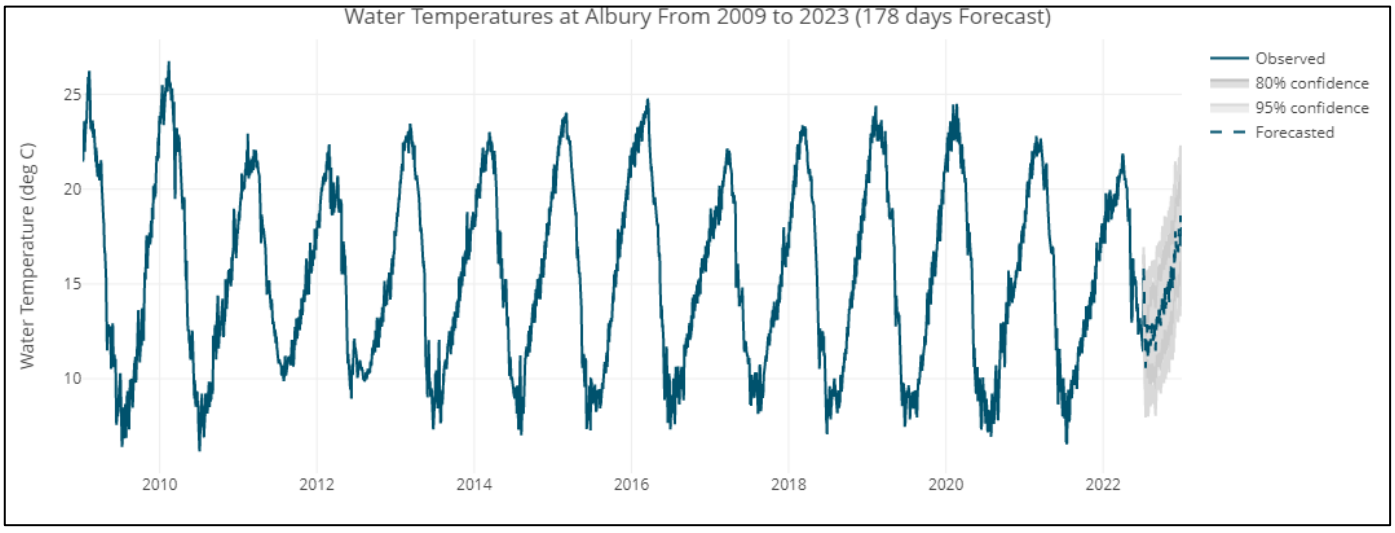

Figure 34: Water Temperatures 178 days Forecast given by the best ARIMA(2,0,2)(0,1,0)[365] (SARIMA) Model

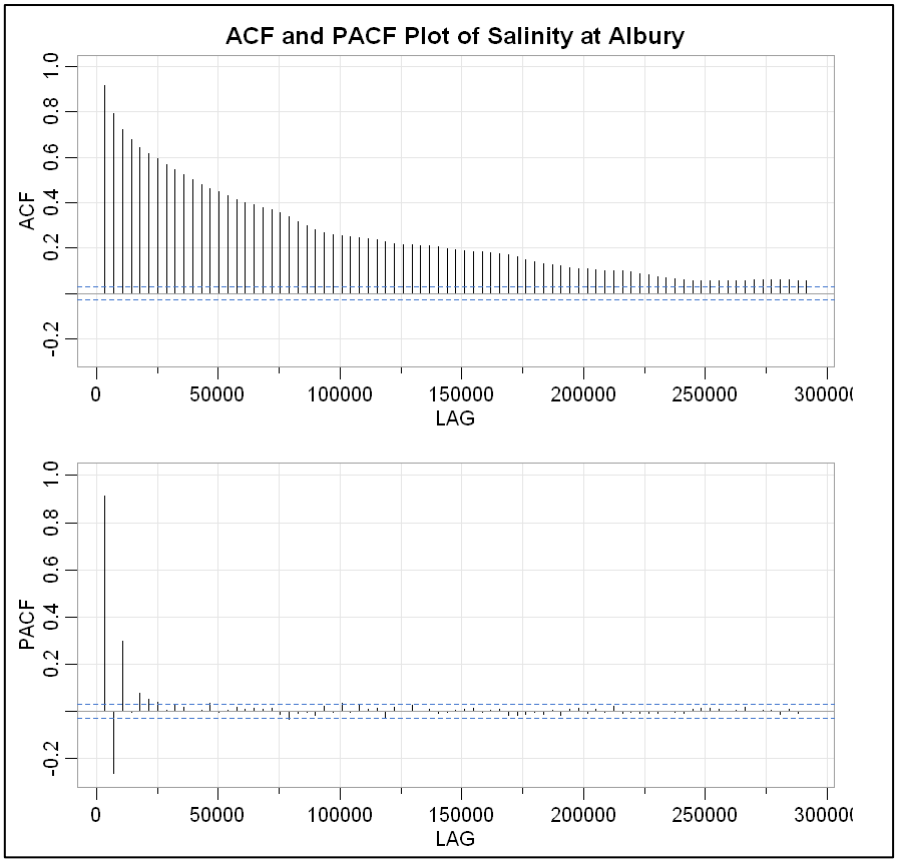

Figure 35: ACF and PACF for Salinity at Albury. Here, ACF display's a seasonal behaviour

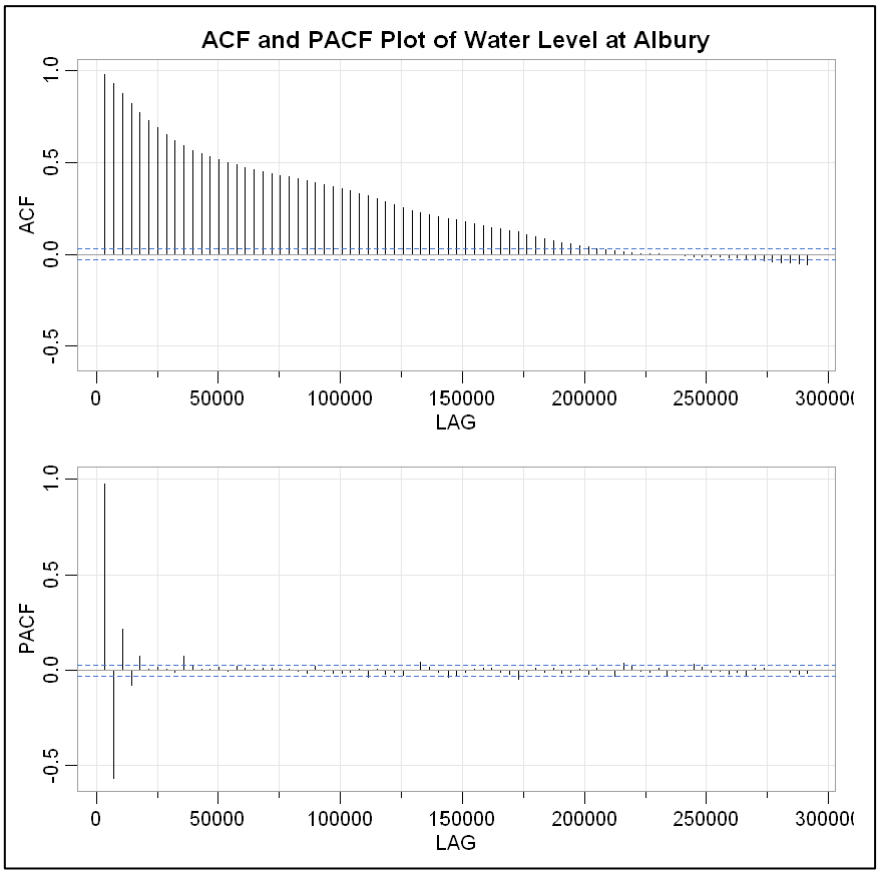

Figure 36: ACF and PACF for Water Level at Albury. Here, ACF display's a seasonal behaviour

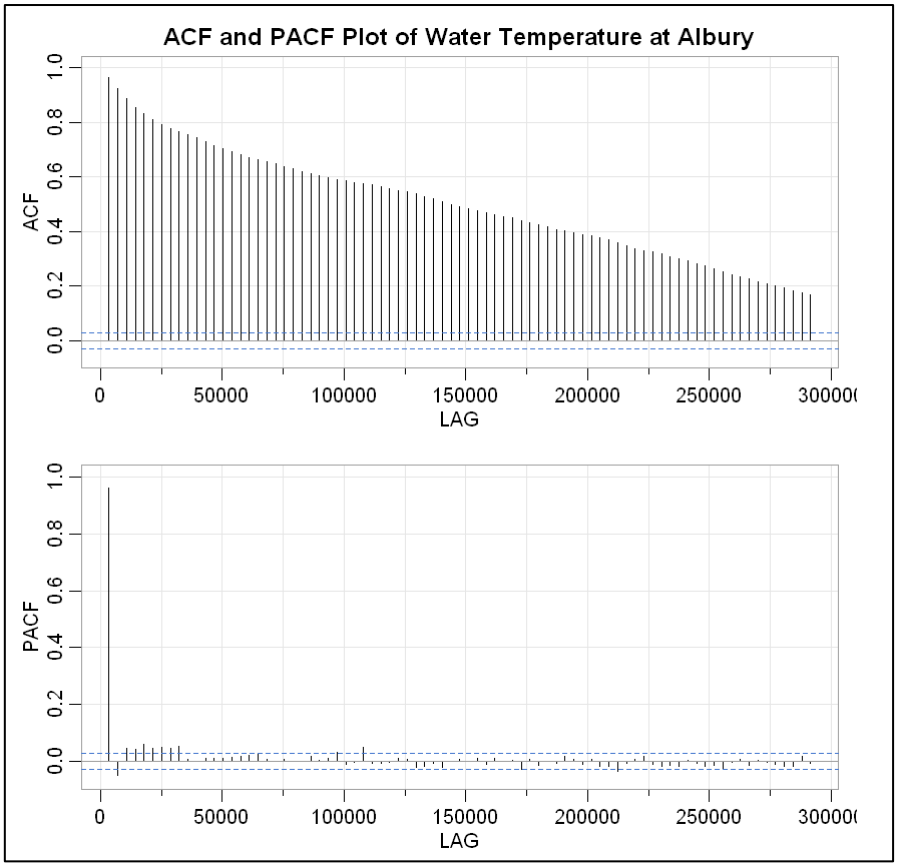

Figure 37: ACF and PACF for Water Temperature at Albury. Here, ACF display's a seasonal behaviour

The table data of Albury Location:

Table 25: Different (Top 5) ARIMA Models Trained for Salinity

ARIMA Models	AIC	AICc	BIC
ARIMA(1,1,2)	24928.43	24928.44	24954.44
ARIMA(0,1,0)	25948.32	25948.32	25954.83
ARIMA(1,1,0)	25720.36	25720.36	25733.37
ARIMA(0,1,1)	25439.77	25439.76	25452.77

Table 26: Different (Top 5) ARIMA Models Trained for Water Levels

ARIMA Models	AIC	AICc	BIC
ARIMA(1,0,1)	-10724.25	-10724.25	-10704.74
ARIMA(1,0,4)	-10920.07	-10920.06	-10881.06
ARIMA(2,0,3)	-10915.79	-10915.78	-10876.78
ARIMA(5,0,0)	-10924.34	-10924.32	-10885.32

Table 27: Different (Top 5) ARIMA Models Trained for Water Temperatures

ARIMA Models	AIC	AICc	BIC
ARIMA(1,0,1)	5103.846	5103.851	5123.353
ARIMA(1,0,4)	5034.9	5034.91	5073.91
ARIMA(2,0,1)	5094.072	5094.081	5120.082
ARIMA(5,0,0)	5051.681	5051.698	5090.695

Table 28: Different (Top 5) SARIMA Models Trained for Salinity

SARIMA Models	AICc
ARIMA(0,0,0)(0,1,0)[365]	35192.01
ARIMA(1,0,0)(0,1,0)[365]	24536.71
ARIMA(2,0,0)(0,1,0)[365]	24258.1
ARIMA(3,0,0)(0,1,0)[365]	23870.35

Table 29: Different (Top 5) SARIMA Models Trained for Water Levels

SARIMA Models	AICc
ARIMA(1,1,1)(0,1,0)[365]	-5112.618
ARIMA(1,1,2)(0,1,0)[365]	-5110.829
ARIMA(2,1,1)(0,1,0)[365]	-5110.046
ARIMA(2,1,2)(0,1,0)[365]	-5236.531

Table 30: Different (Top 5) SARIMA Models Trained for Water Temperatures

SARIMA Models	AICc
ARIMA(1,0,1)(0,1,0)[365]	7635.622
ARIMA(1,0,2)(0,1,0)[365]	7618.019
ARIMA(2,0,1)(0,1,0)[365]	7627.981
ARIMA(2,0,2)(0,1,0)[365]	7544.7

The plot of Biggara Location:

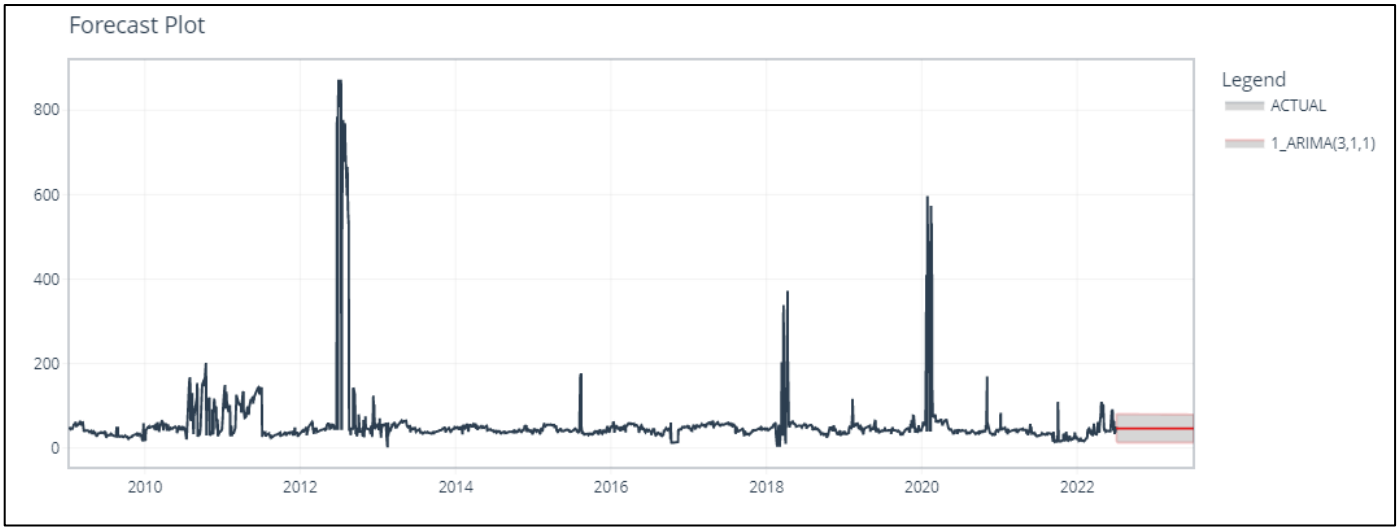

Figure 38: Salinity Levels 365 days Forecast given by the best ARIMA(3,1,1) Model

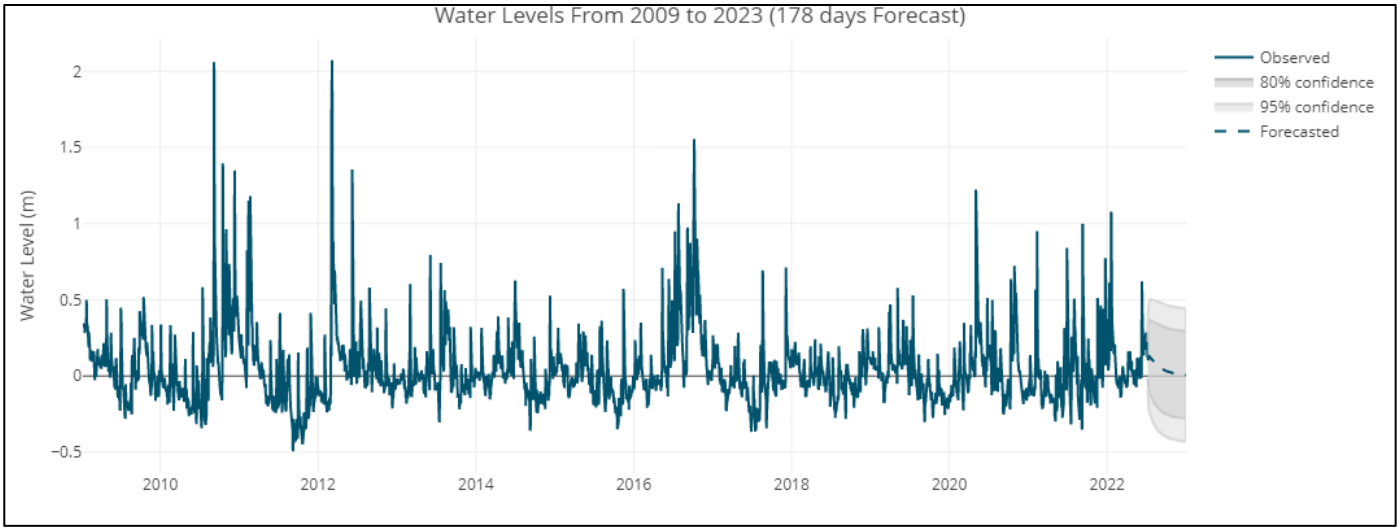

Figure 39: Water Levels 178 days Forecast given by the best ARIMA(3,0,2) Model

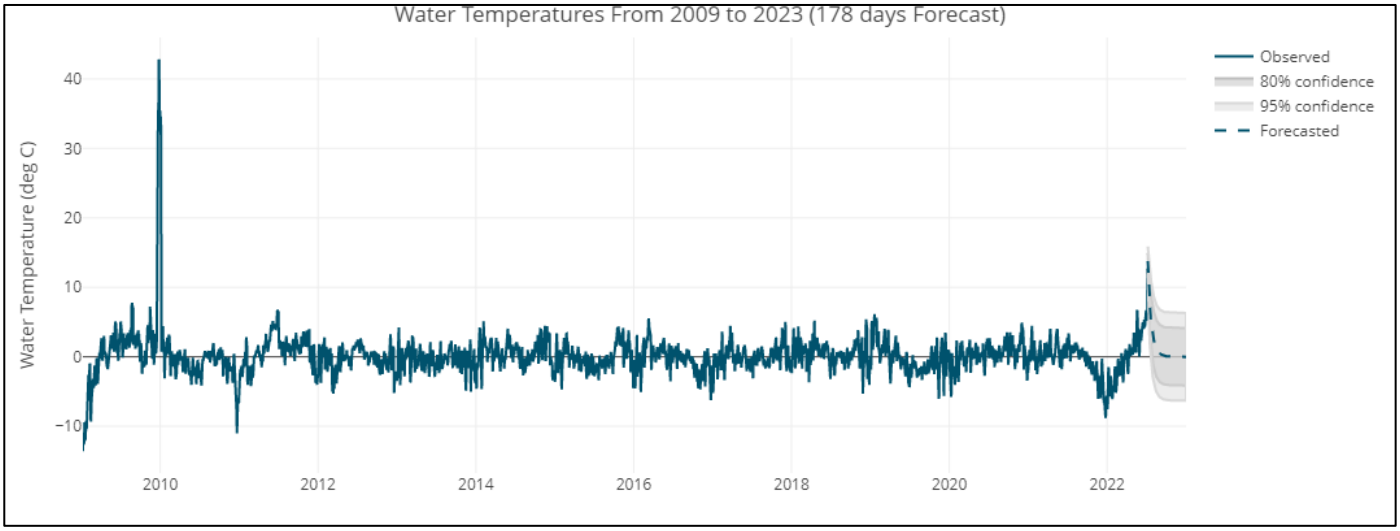

Figure 40: Water Temperatures 178 days Forecast given by the best ARIMA(1,0,4) Model

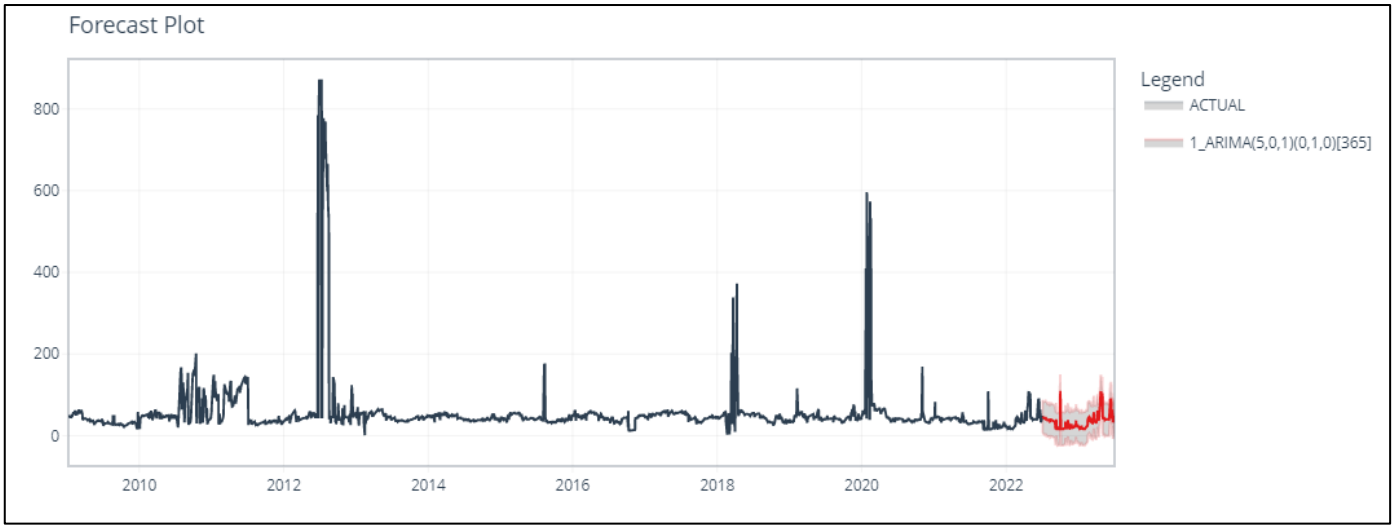

Figure 41: Salinity Levels 365 days Forecast given by the best ARIMA(5,0,1)(0,1,0)[365] (SARIMA) Model

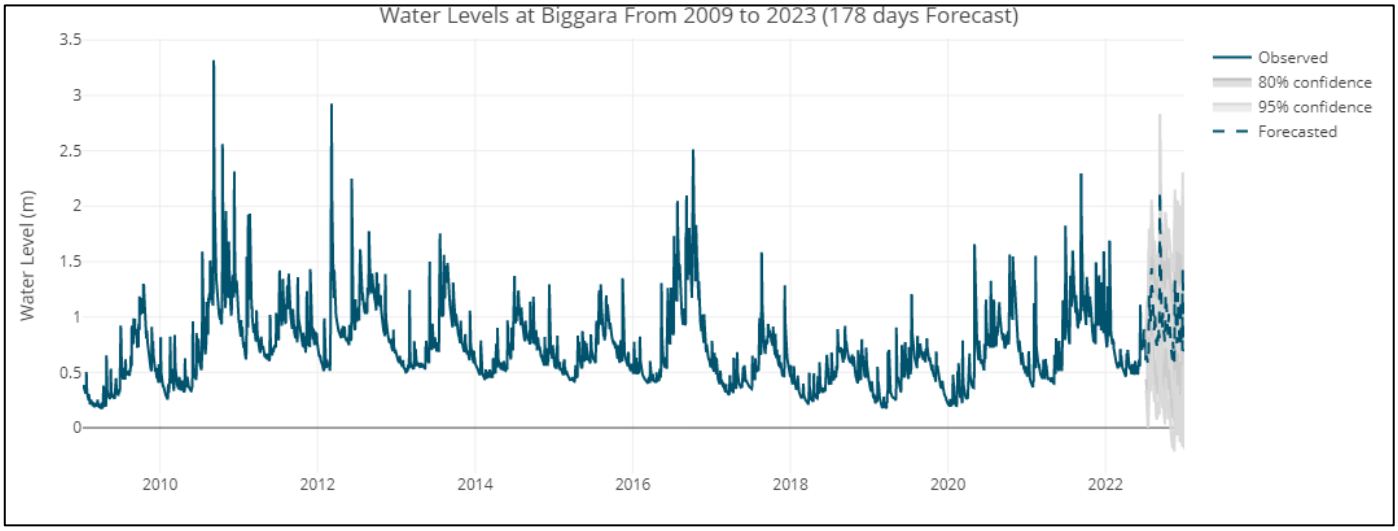

Figure 42: Water Levels 178 days Forecast given by the best ARIMA(2,1,2)(0,1,0)[365] Model

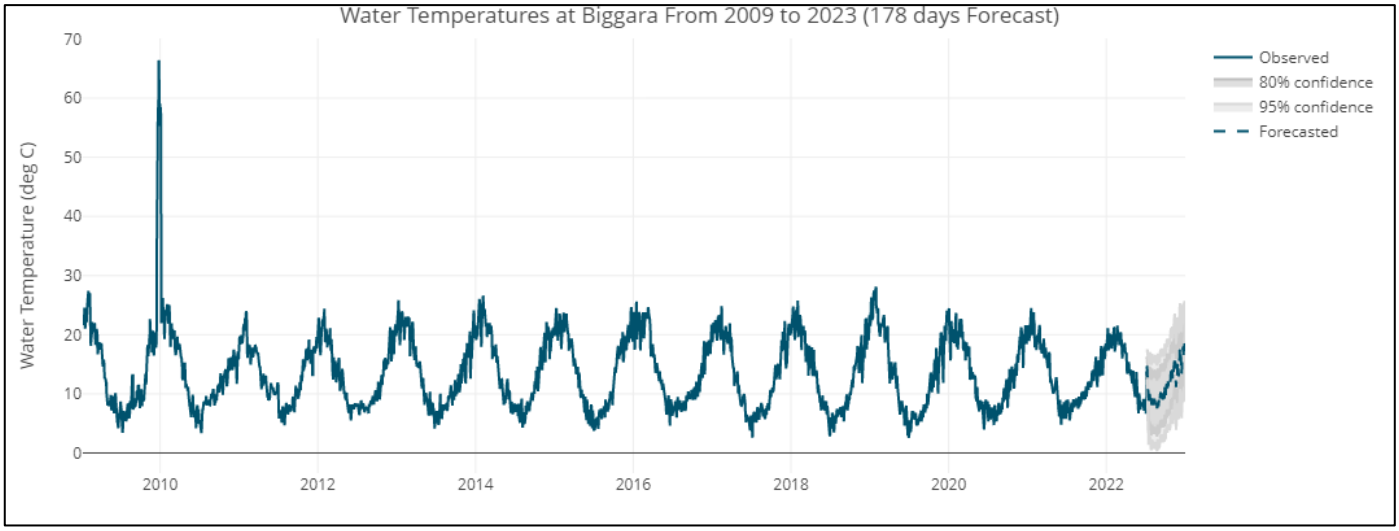

Figure 43: Water Temperatures 178 days Forecast given by the best ARIMA(1,0,2)(0,1,0)[365] Model

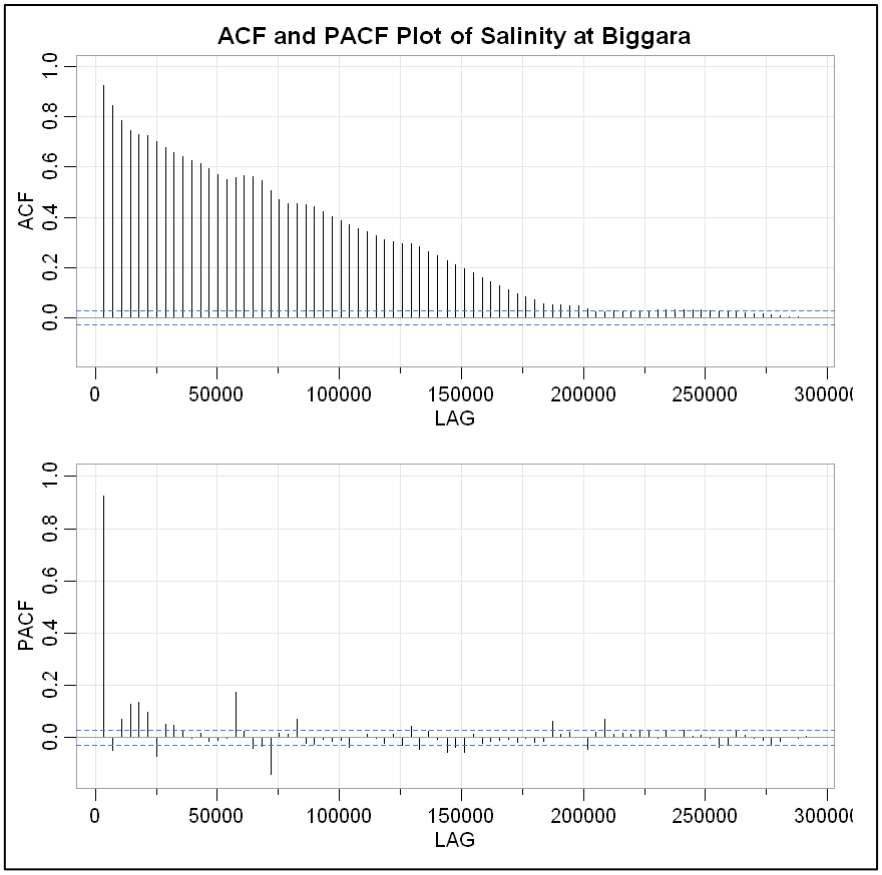

Figure 44: ACF and PACF for Salinity at Biggara. Here, ACF display's a seasonal behaviour

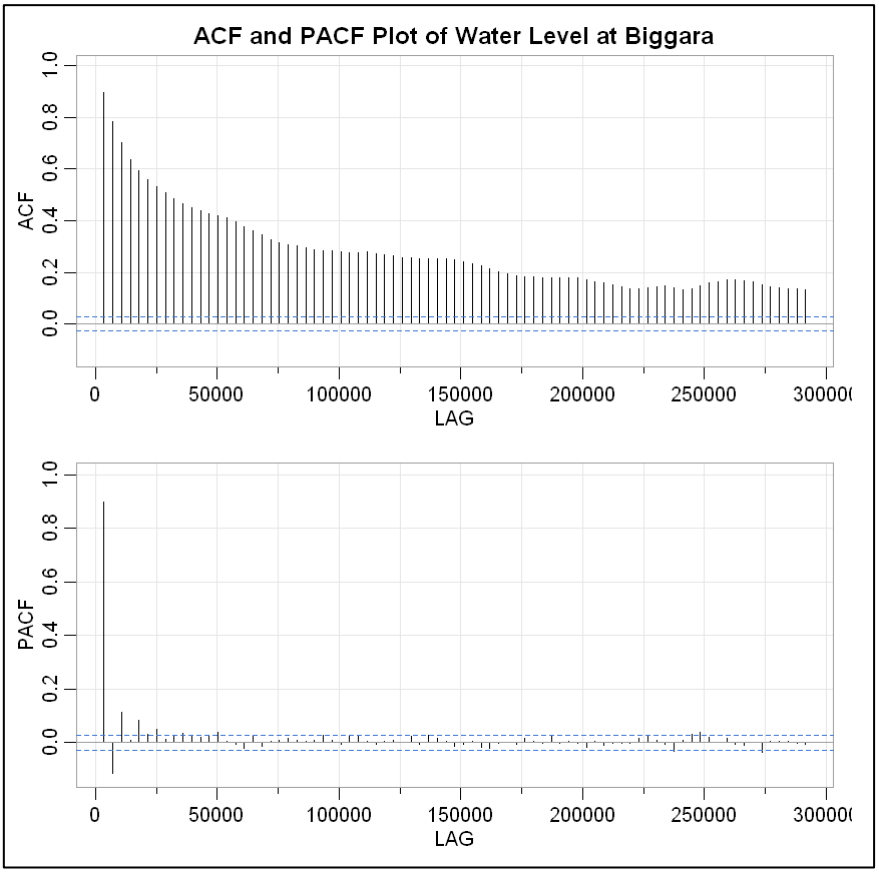

Figure 45: ACF and PACF for Water Level Biggara. Here, ACF display's a seasonal behaviour

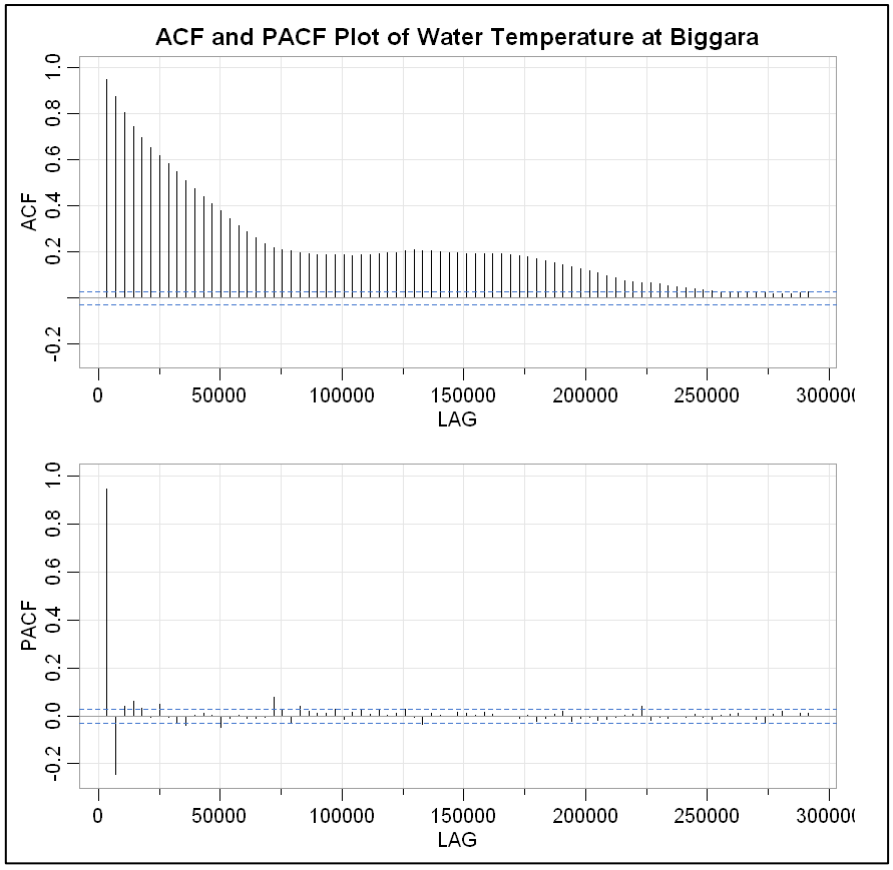

Figure 46: ACF and PACF for Water Temperature Biggara. Here, ACF display's a seasonal behaviour

The table data of Biggara Location:

Table 31: Different (Top 5) ARIMA Models Trained for Salinity

ARIMA Models	AIC	AICc	BIC
ARIMA(2,1,1)	47001.56	47001.57	47027.57
ARIMA(2,1,0)	47271.69	47271.7	47291.2
ARIMA(3,1,1)	46983.66	46983.67	47016.17
ARIMA(3,1,0)	47150.99	47150.99	47176.99

Table 32: Different (Top 5) ARIMA Models Trained for Water Levels

ARIMA Models	AIC	AICc	BIC
ARIMA(1,0,1)	-9142.337	-9142.332	-9122.83
ARIMA(1,0,4)	-9235.177	-9235.16	-9196.164
ARIMA(3,0,2)	-9258.58	-9258.57	-9219.57
ARIMA(4,0,1)	-9189.403	-9189.386	-9150.389

Table 33: Different (Top 5) ARIMA Models Trained for Water Temperatures

ARIMA Models	AIC	AICc	BIC
ARIMA(1,0,1)	13179.3	13179.31	13198.81
ARIMA(1,0,4)	13137.73	13137.75	13176.75
ARIMA(3,0,2)	13142.11	13142.12	13181.12
ARIMA(4,0,1)	13142.06	13142.08	13181.07

Table 34: Different (Top 5) SARIMA Models Trained for Salinity

SARIMA Models	AICc
ARIMA(3,0,0)(0,1,0)[365]	43715.19
ARIMA(4,0,0)(0,1,0)[365]	43646.26
ARIMA(5,0,0)(0,1,0)[365]	43566.49
ARIMA(5,0,1)(0,1,0)[365]	43544.85

Table 35: Different (Top 5) SARIMA Models Trained for Water Levels

SARIMA Models	AICc
ARIMA(1,1,1)(0,1,0)[365]	-3197.952
ARIMA(1,1,2)(0,1,0)[365]	-3467.669
ARIMA(2,1,1)(0,1,0)[365]	-3452.789
ARIMA(2,1,2)(0,1,0)[365]	-3494.824

Table 36: Different (Top 5) SARIMA Models Trained for Water Temperature

SARIMA Models	AICc
ARIMA(1,0,1)(0,1,0)[365]	13688.92
ARIMA(1,0,2)(0,1,0)[365]	13665.81
ARIMA(2,0,1)(0,1,0)[365]	13673.46
ARIMA(2,0,2)(0,1,0)[365]	13667.57

APPENDIX - B

The plot of Murray Bridge Location:

Figure 47: Salinity Levels Forecast from May 2021 to July 2022 (Test Dataset) given by the best SVM Model (rbf_sigma=0.75, cost=0.1, margin=0.01)

Figure 48: Salinity Levels 365 days Forecast given by the best SVM Model (rbf_sigma=0.75, cost=0.1, margin=0.01)

The table data of Murray Bridge Location:

Table 37: Different (Top 5) SVM Models with different Hyperparameters Cross-Validated for Salinity

cost	rbf_sigma	margin	RMSE	SD
0.1	0.75	0.04	78.650	1.073
0.1	0.75	0.05	78.652	1.061
0.1	0.75	0.03	78.671	1.075
0.1	0.75	0.01	78.683	1.058
0.1	0.75	0.02	78.701	1.066

Table 38: Low correlation between Salinity, Water Level and Water Temperature. Therefore, we can use Water Level and Water Level as predictor variables for Salinity

	Water Level	Salinity	Water Temperature
Water Level	1.000	-0.260	-0.002
Salinity	-0.260	1.000	-0.064
Water Temperature	-0.002	-0.064	1.000

The plot of Colignan Location:

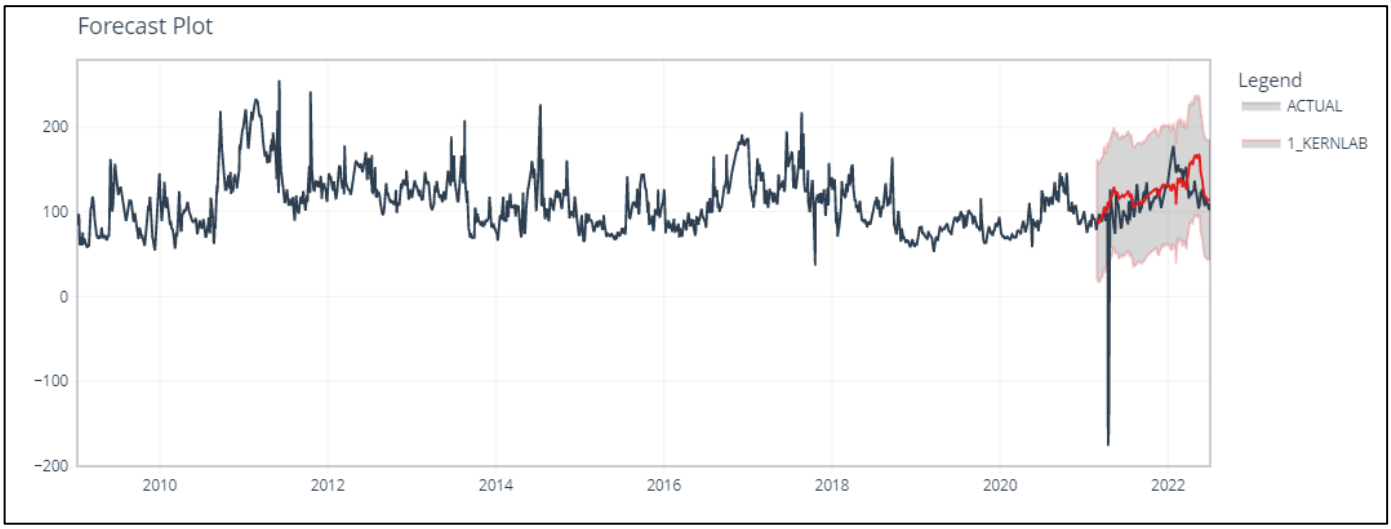

Figure 49: Salinity Levels Forecast from May 2021 to July 2022 (Test Dataset) given by the best SVM Model (rbf_sigma=1, cost=0.31)

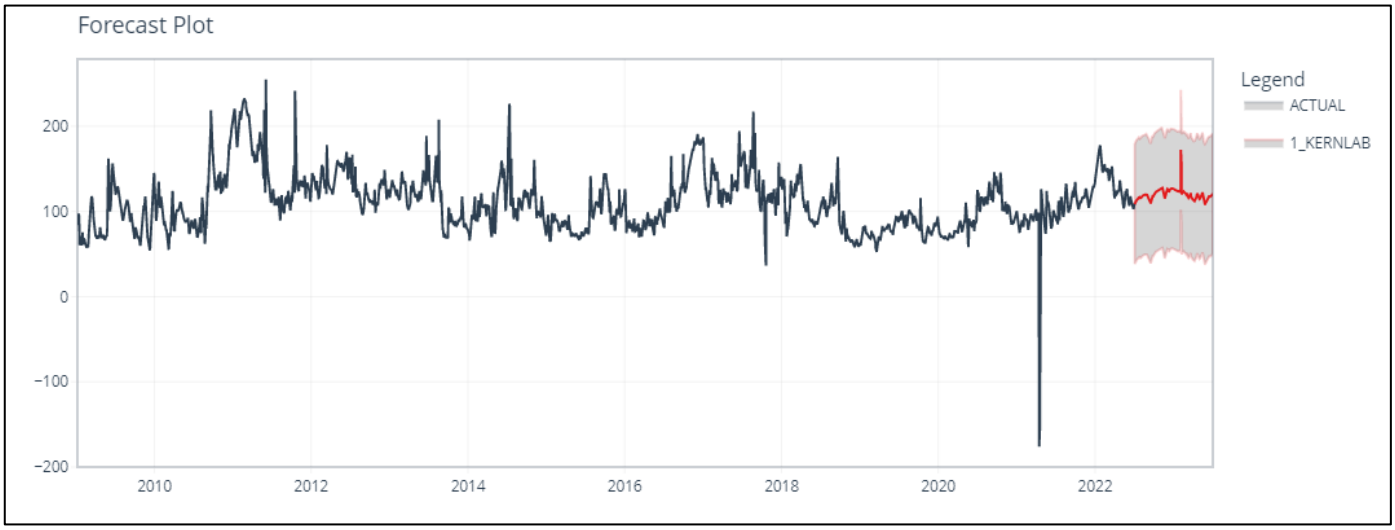

Figure 50: Salinity Levels 365 days Forecast given by the best SVM Model (rbf_sigma=1, cost=0.31)

The table data of Colignan Location:

Table 39: Different (Top 5) SVM Models with different Hyperparameters Cross-Validated for Salinity

cost	rbf_sigma	RMSE	SD
32.00	1	15.60	0.3139
10.08	1	16.12	0.3215
3.17	1	16.68	0.3212
1.00	1	17.53	0.3464
0.31	1	19.07	0.3656

Table 40: Low correlation between Salinity, Water Level and Water Temperature. Therefore, we can use Water Level and Water Level as predictor variables for Salinity

	Water Level	Salinity	Water Temperature
Water Level	1.00	0.35	-0.13
Salinity	0.35	1.00	-0.06
Water Temperature	-0.13	-0.06	1.00

The plot of Albury Location:

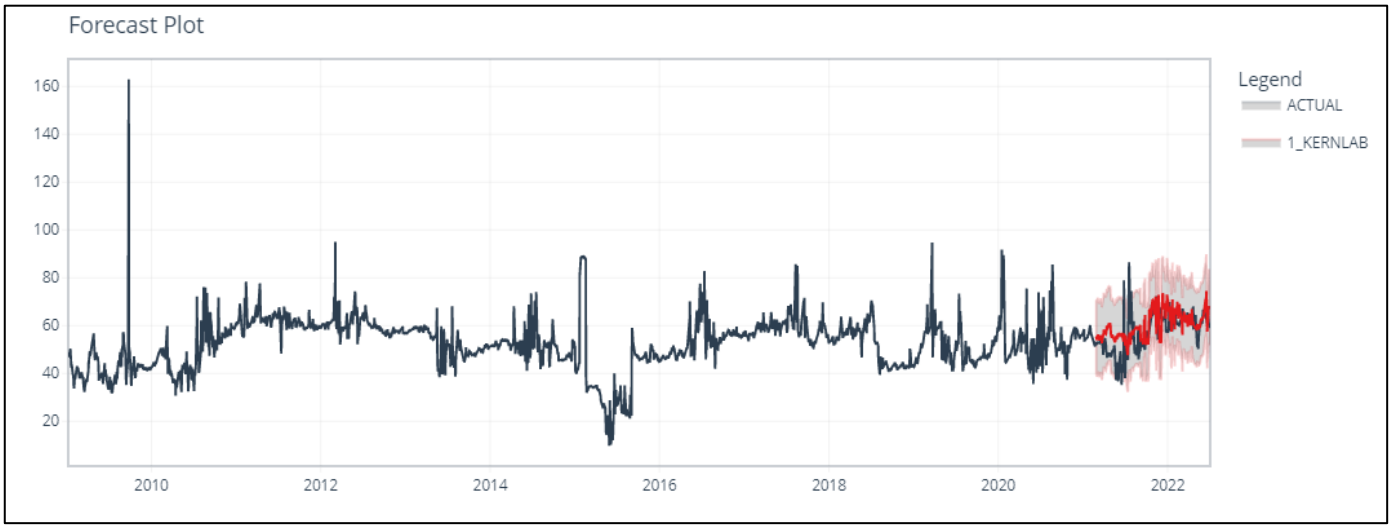

Figure 51: Salinity Levels Forecast from May 2021 to July 2022 (Test Dataset) given by the best SVM Model (rbf_sigma=1, cost=1)

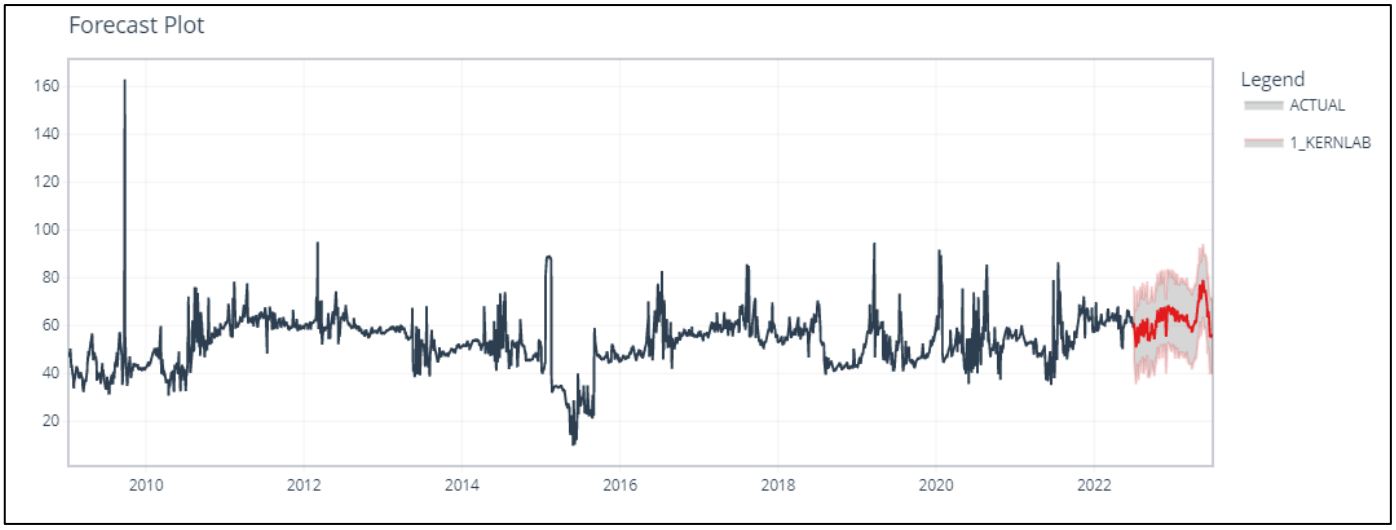

Figure 52: Salinity Levels 365 days Forecast given by the best SVM Model (rbf_sigma=1, cost=1)

The table data of Albury Location:

Table 41: Different (Top 5) SVM Models with different Hyperparameters Cross-Validated for Salinity

cost	rbf_sigma	RMSE	SD
32.00	1	7.43	0.39
10.08	1	7.65	0.37
3.17	1	7.83	0.35
1.00	1	7.96	0.32
0.31	1	8.19	0.31

Table 42: Low correlation between Salinity, Water Level and Water Temperature. Therefore, we can use Water Level and Water Level as predictor variables for Salinity

	Water Level	Salinity	Water Temperature
Water Level	1.00	0.16	0.44
Salinity	0.16	1.00	0.06
Water Temperature	0.44	0.06	1.00

The plot of Biggara Location:

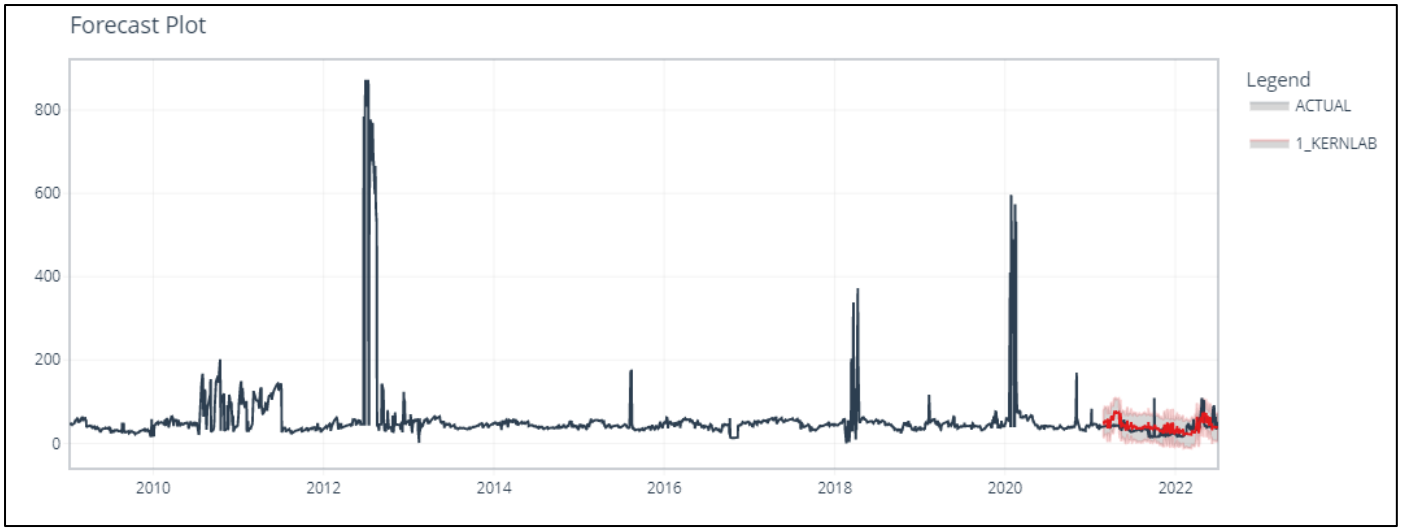

Figure 53: Salinity Levels Forecast from May 2021 to July 2022 (Test Dataset) given by the best SVM Model (rbf_sigma=1, cost= 3.17)

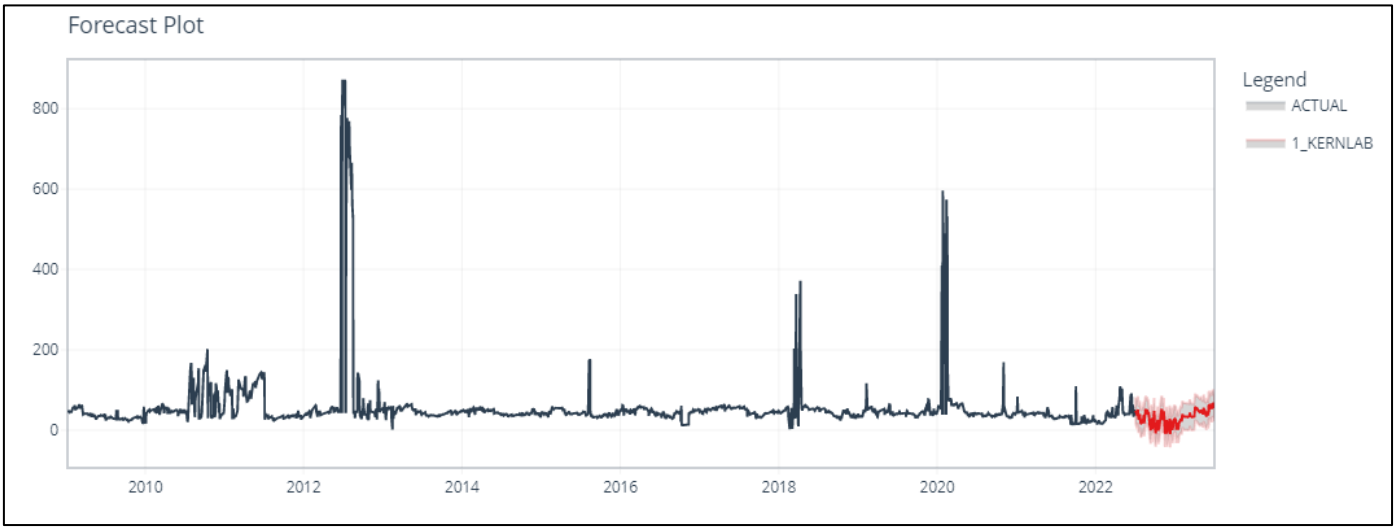

Figure 54: Salinity Levels 365 days Forecast given by the best SVM Model (rbf_sigma=1, cost= 3.17)

The table data of Biggara Location:

Table 43: Different (Top 5) SVM Models with different Hyperparameters Cross-Validated for Salinity

cost	rbf_sigma	RMSE	SD
32.00	1	76.81	5.399
10.08	1	77.44	5.397
3.17	1	77.85	5.350
1.00	1	78.27	5.295
0.31	1	78.66	5.238

Table 44: Low correlation between Salinity, Water Level and Water Temperature. Therefore, we can use Water Level and Water Level as predictor variables for Salinity

	Water Level	Salinity	Water Temperature
Water Level	1.00	-0.37	-0.39
Salinity	-0.37	1.00	0.17
Water Temperature	-0.39	0.17	1.00

APPENDIX - C

The hyperparameter optimization plot data of Murray Bridge Location:

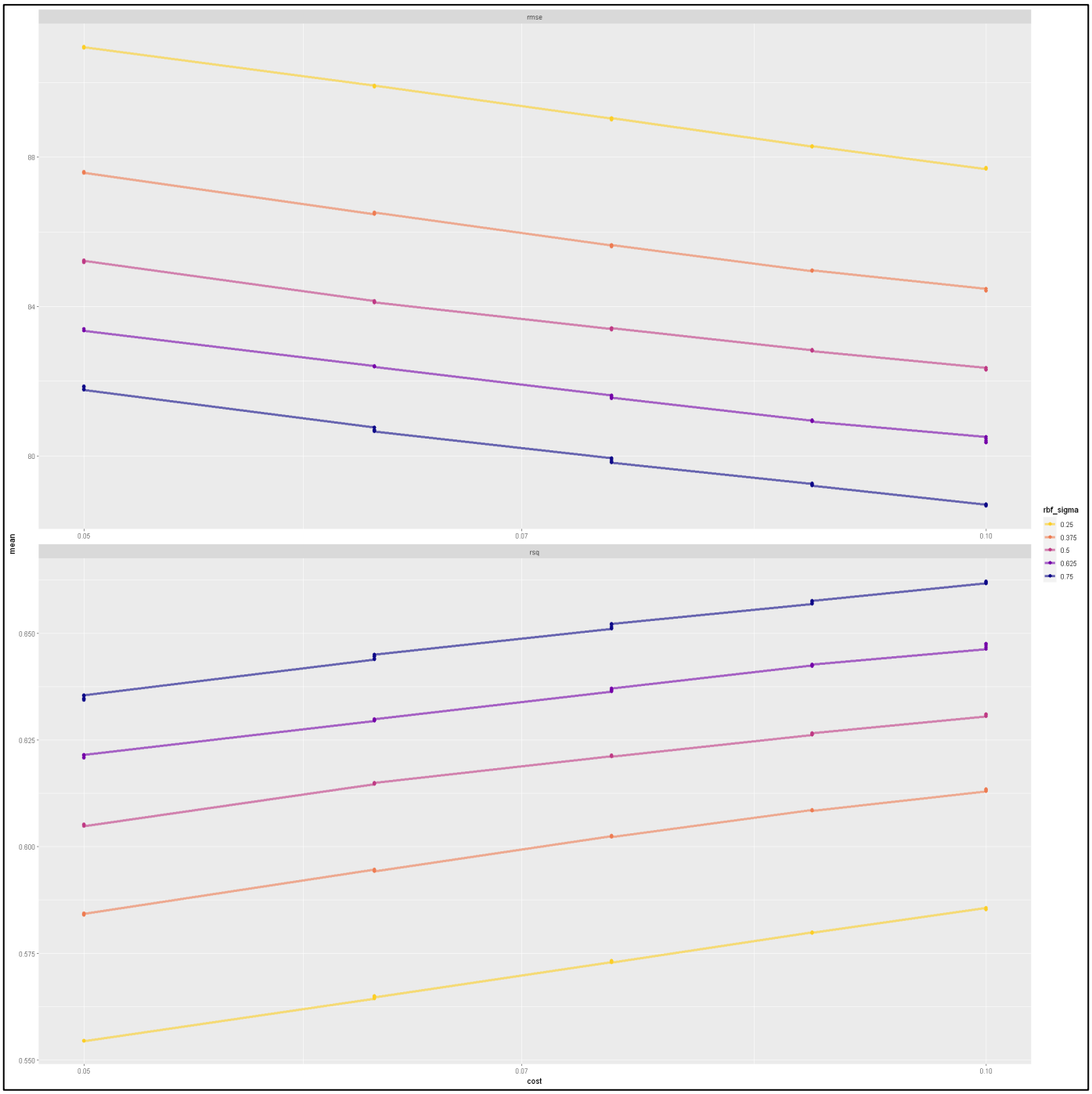

Figure 55: Different Hyperparameter values being Cross-Validated for choosing the best SVM Model (rbf_sigma=0.75, cost=0.1, margin=0.01). The top 5 Hyperparameter values are provided in Table 37

The hyperparameter optimization plot data of Colignan Location:

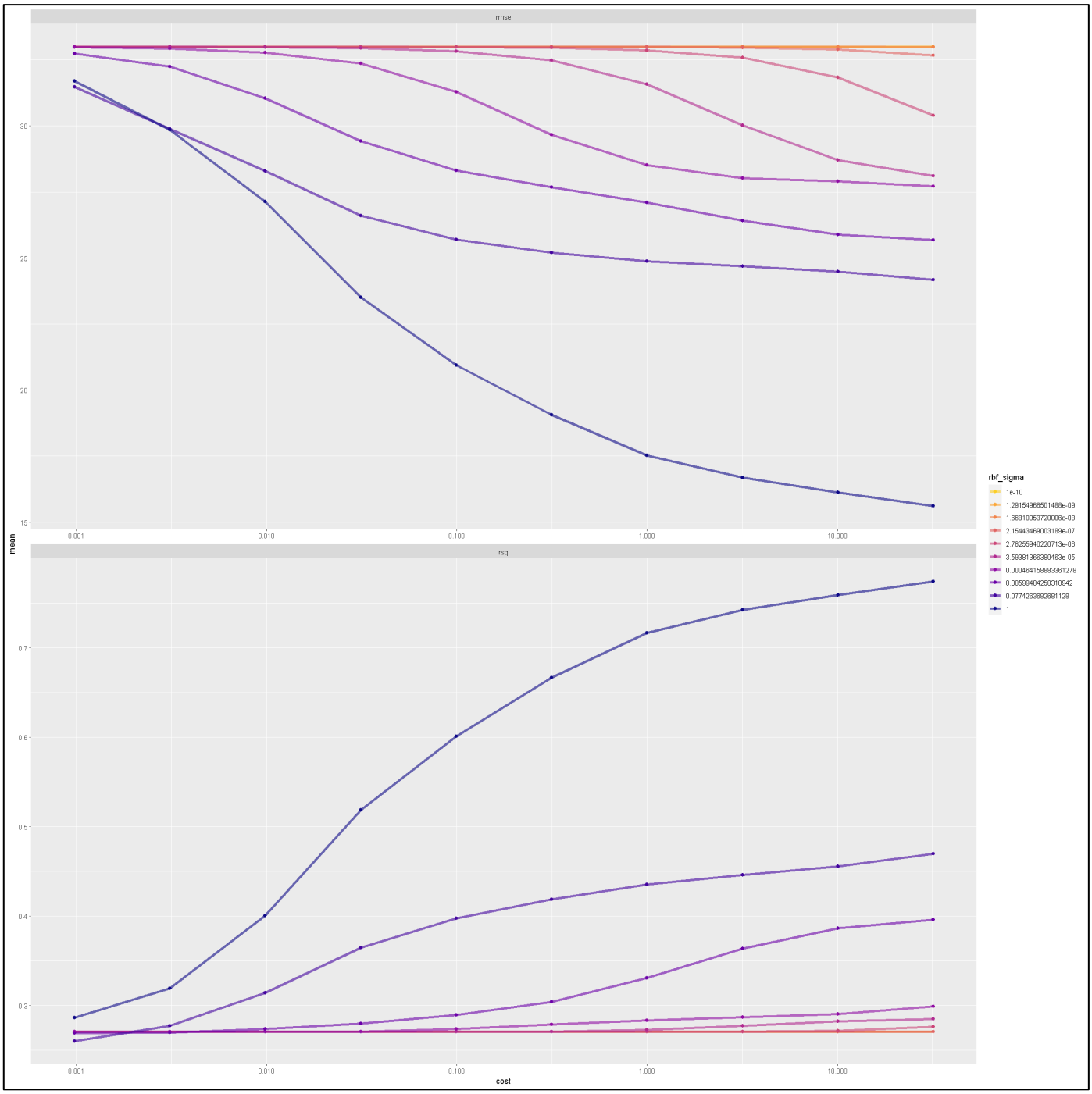

Figure 56: Different Hyperparameter values being Cross-Validated for choosing the best SVM Model (rbf_sigma=1, cost=0.31). The top 5 Hyperparameter values are provided in Table 39

The hyperparameter optimization plot data of Albury Location:

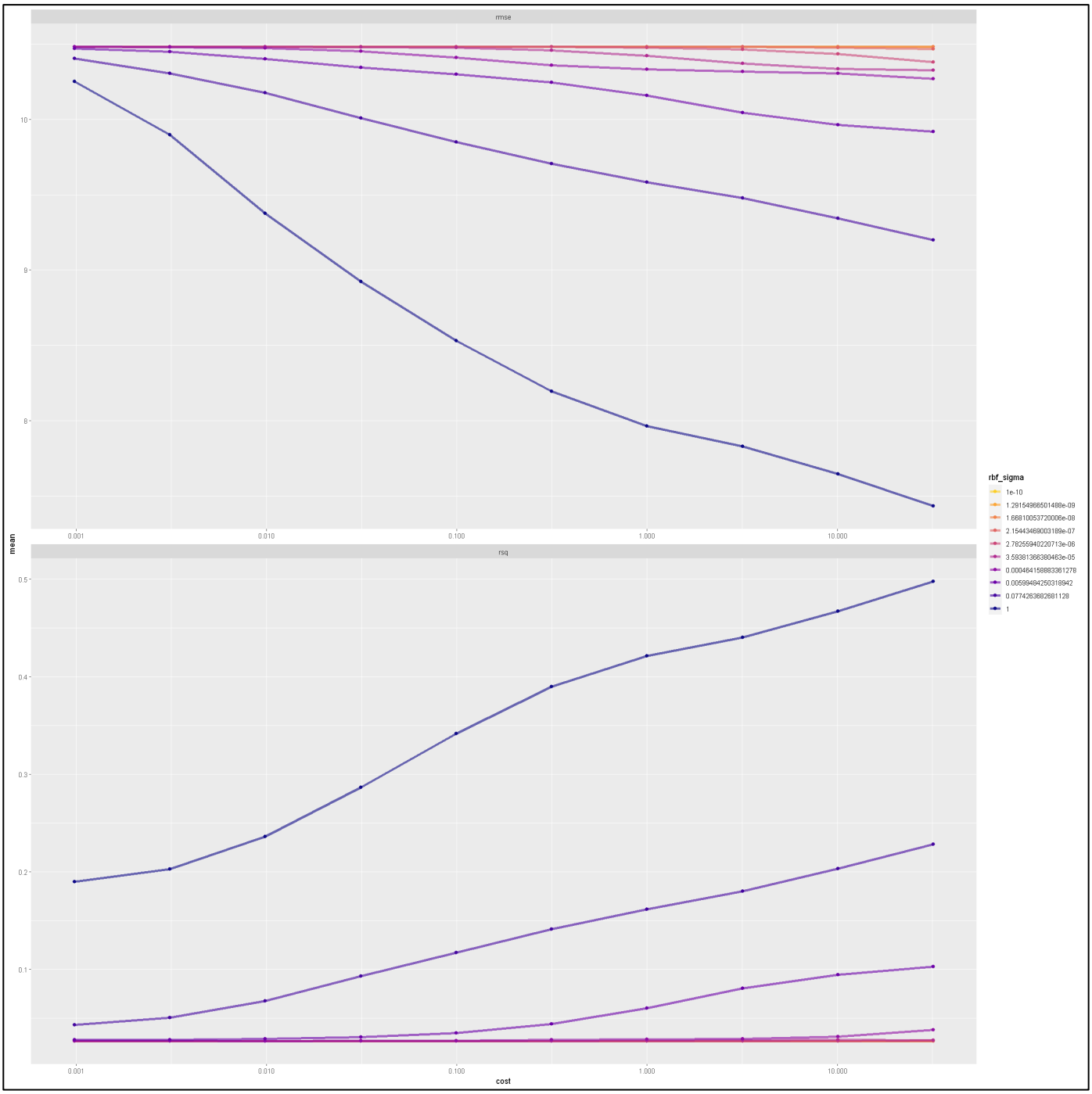

Figure 57: Different Hyperparameter values being Cross-Validated for choosing the best SVM Model (rbf_sigma=1, cost=1). The top 5 Hyperparameter values are provided in Table 41

The hyperparameter optimization plot data of Biggara Location:

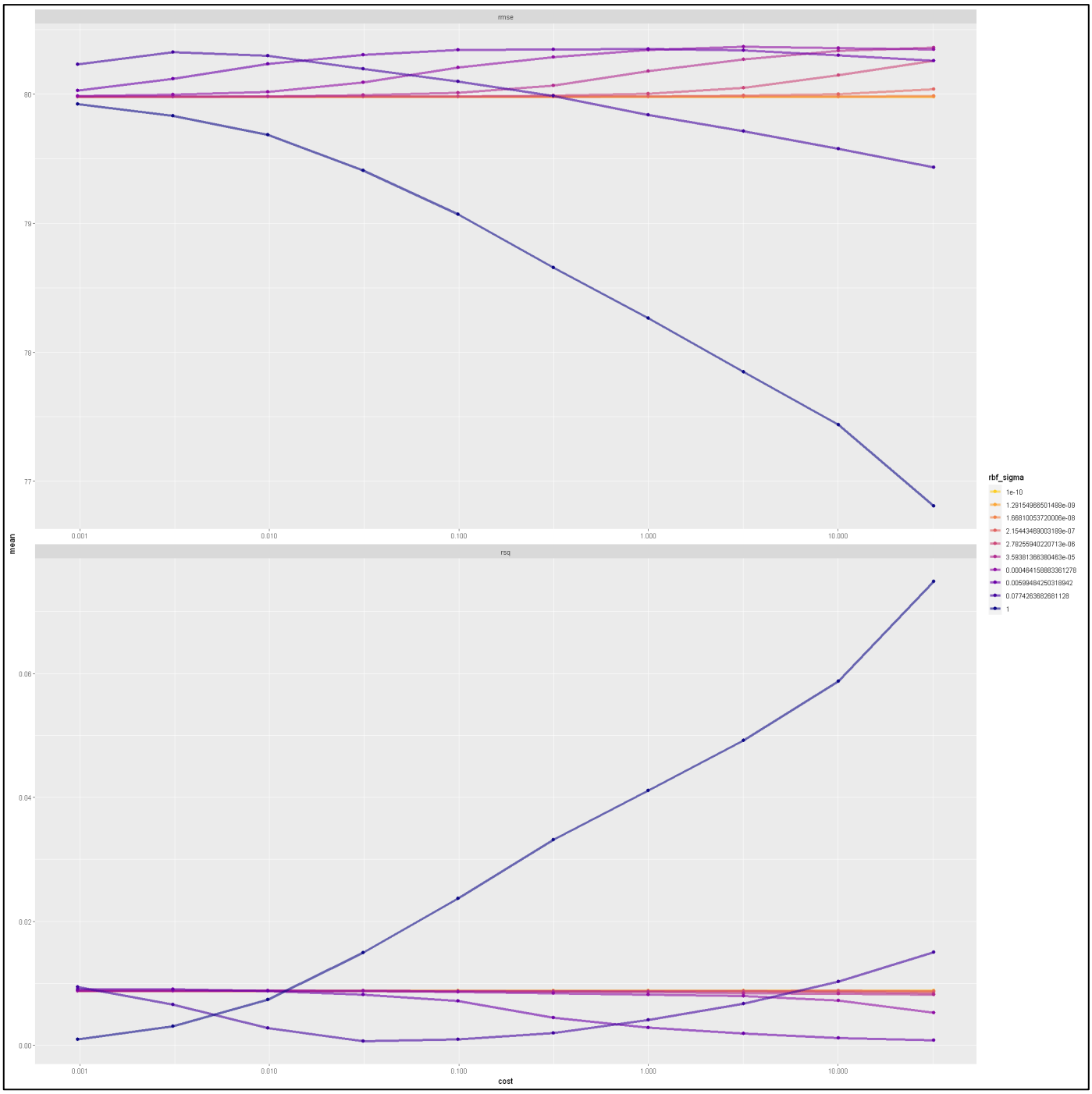

Figure 58: Different Hyperparameter values being Cross-Validated for choosing the best SVM Model (rbf_sigma=1, cost=3.17). The top 5 Hyperparameter values are provided in Table 43

APPENDIX – D

Data Parameter and Location Selection

The MDBA Authority Data, as explained above, has three key data types or variables:

1) Electrical Conductivity (Salinity):

   The Electrical Conductivity (E.C) is measured in micro-siemens per centimetre, and it gives an indication of the amount of salt in the river. There is a strong correlation between the salinity of the river and the E.C. measure. The advantage of measuring salinity in form of E.C. is that E.C. can be documented in real-time or continuously (MDBA, 2022a).

2) Gauge Level (Water Level):

   The Gauge level measure is utilized to represent the current river water level on a gauge board. These water levels are further utilized to predict the river water flow and reservoir storage volume. The Gauge Level is measured in Australian Height Datum meters (AHD). The Meters AHD is the measurement of ascent from a mean sea level (mean sea level is taken as zero) (MDBA, 2022a).

3) Water Temperature:

   The Water Temperature is measured in terms of the degree Celsius, and it varies both seasonally and daily. The dependency of water temperature is simultaneously on both water discharge from reservoirs and surrounding air temperature (MDBA, 2022a).

Data Decomposition

There are two types of data decomposition methods:

1) Additive Decomposition:

   In Additive Decomposition, the components are broken in such a way that the original time series can be formed by just adding up the Seasonal, Trend-Cycle and Remainder components. The additive decomposition is best suitable when the fluctuation around the trend cycle or seasonal pattern seems to be non-proportional to the level of the time series (Rob & George, 2018a). The formula of additive decomposition is given below in Figure 59.

$$y_t = S_t + T_t + R_t$$

Figure 59: The formula for Additive Decomposition

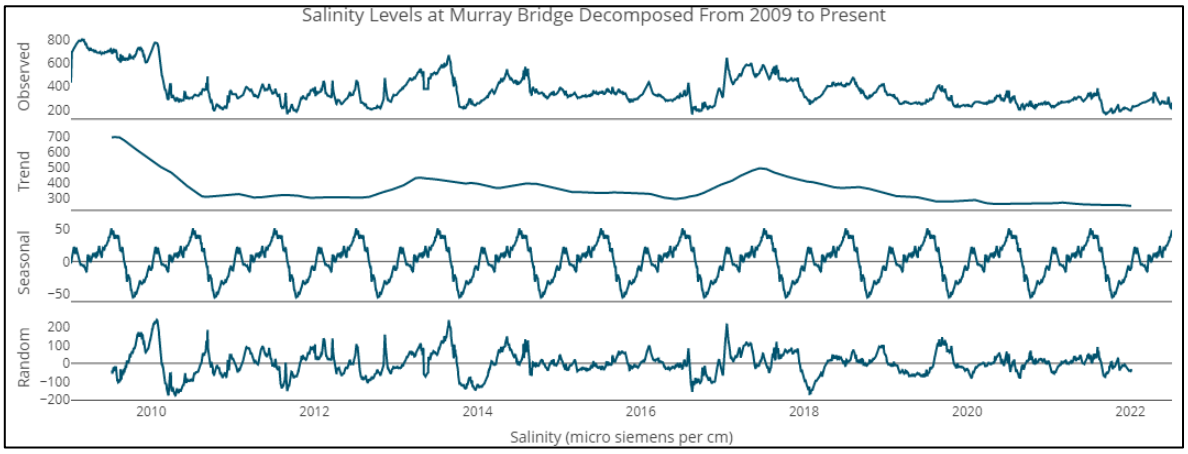

Figure 4: Decomposition of Salinity Time Series into its seasonal, trend and random components at Murray Bridge

2) Multiplicative Decomposition:

In Multiplicative Decomposition, the components are broken in such a way that the original time series can be formed by just multiplying the Seasonal, Trend-Cycle and Remainder components. The multiplicative decomposition is best suitable when the fluctuation around the trend cycle or seasonal pattern seems to vary with the level of the time series (Rob & George, 2018a). The formula of multiplicative decomposition is given below in Figure 60.

$$y_t = S_t \times T_t \times R_t$$

Figure 60: The formula for Multiplicative Decomposition

Box-Cox Transformation

The logarithm being used in the Box-Cox Transformation always has a base e (i.e., natural logarithm). Therefore, if the value of $\lambda = 0$, then natural logarithm transformations are utilized and if the value of $\lambda \neq 0$, power transformations are utilized with $1/\lambda$ as the scaling factor. Moreover, if $\lambda = 1$, then the Formula in Figure 3 becomes $w_t = y_t - 1$, thus, the times series data is just shifted downwards without any change in the original time series (Rob & George, 2018a).

The Box-Cox Transformation is performed with the help of the BoxCox() method of forecast Library in R Programming Language (Hyndman et. al, 2022). The BoxCox() method utilizes a slightly modified version of the Box-Cox Transformation, which is explained in Bickel & Doksum (1981), this enables the use of negative values of y_t when $\lambda > 0$. The formula is given in Figure 61.

$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ \text{sign}(y_t)(|y_t|^{\lambda} - 1)/\lambda & \text{otherwise.} \end{cases}$$

Figure 61: Formula of modified Box-Cox Transform where $1/\lambda$ is the scaling factor

The Inverse of the Box-Cox Transformation can also be performed to get the values of the original scale time series (Rob & George, 2018a). The formula is provided in Figure 62.

$$y_t = \begin{cases} \exp(w_t) & \text{if } \lambda = 0; \\ \text{sign}(\lambda w_t + 1)|\lambda w_t + 1|^{1/\lambda} & \text{otherwise.} \end{cases}$$

Figure 62: Formula of Inverse of Box-Cox Transform

The Box-Cox Transform reduces the variance in time series data which in turn helps in making the series more stationary.

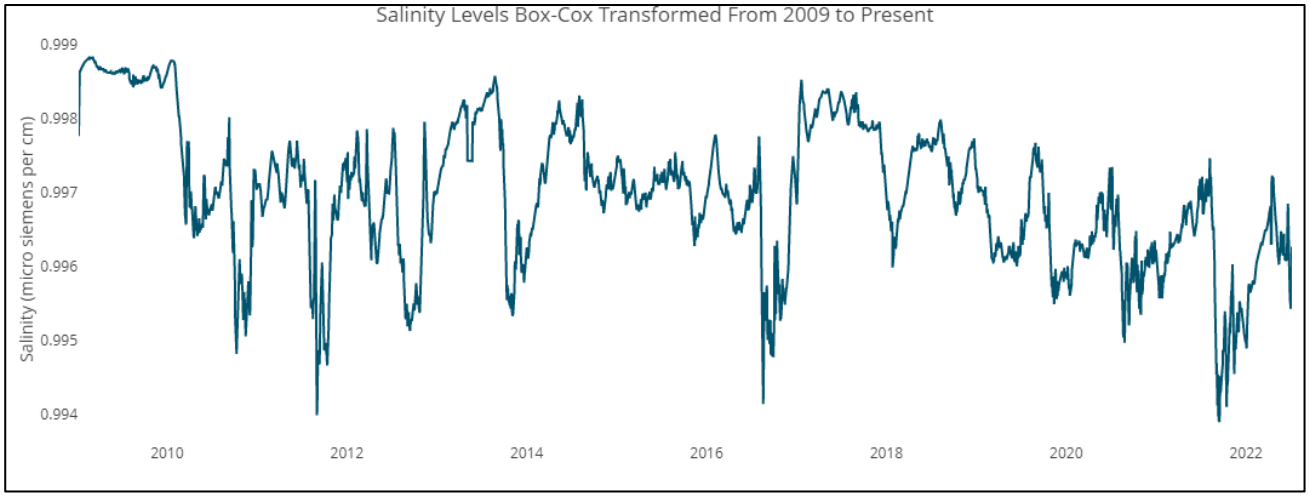

Figure 63: Box-Cox Transformation of Salinity Time-Series makes the variance of the data constant at Murray Bridge

Data Stationarity Tests

To overcome the problem of non-stationarity in time series data, several Stationary Tests are utilized, which are explained below:

1) Augmented Dickey-Fuller test (ADF Test):

   The Augmented Dickey-Fuller test is an Autoregressive Unit Root Test that is utilized to measure the presence of unit roots in the time series data. The Null Hypothesis of the ADF Test is that the time series has a unit root. In contrast, the Alternative Hypothesis is that the time series has no unit root. Hence, if the null hypothesis is rejected, then there is strong evidence that the time series is non-stationary (Statsmodels, 2022; Eric, 2014).

   The ADF test is done by utilizing the adf.test function of the aTSA Library in R Programming Language (Debin, 2015).

2) Kwiatkowski-Phillips-Schmidt-Shin (KPSS):

   The Kwiatkowski-Phillips-Schmidt-Shin test is a Stationarity Test that is utilized to check whether the time series data is stationary or not. The Null Hypothesis of the KPSS Test is that the time series has a stagnant trend. While the Alternative Hypothesis is that the time series has a unit root. Hence, if the null hypothesis is rejected, then there is strong evidence that the time series is non-stationary and has a unit root which is the opposite when compared to ADF (Statsmodels, 2022; Piotr & Gabriel, 2016).

   The KPSS test is done by utilizing the kpss.test function of the aTSA Library in R Programming Language (Debin, 2015).

ARIMA Model

ACF, PACF, Box-Ljung test, AIC, AICc and BIC are all explained below:

1) Auto-Correlation Function (ACF) plot:

   Correlation measures the amount of linear relationship between 2 variables. Similarly, autocorrelation measures the extent of the linear relationship between the original time series and its lagged version (Rob & George, 2018d). The formula for autocorrelation coefficients is given in Figure 64.

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2},$$

Figure 64: Formula of autocorrelation coefficients where r_k is the autocorrelation of k times lagged time series

Here, r_k is the autocorrelation of k times lagged time series, and T is the length of the time series. The plot of these autocorrelation coefficients shows the Auto Correlation Function (ACF) (Rob & George, 2018d).

ARIMA(p,d,0) model is selected when the ACF of the time series is sinusoidal or exponentially dampening, and there is a measurable spike at lag p in the plot of PACF, however none after the lag p (Rob, 2017).

2) Partial Auto-Correlation Function (PACF) plot:

Partial Autocorrelation is a conditional autocorrelation in which the relationship between y_t and y_{t-k} is measured by removing the effects of other time lags – 1, 2, 3, …, k-1 (Rob, 2017).

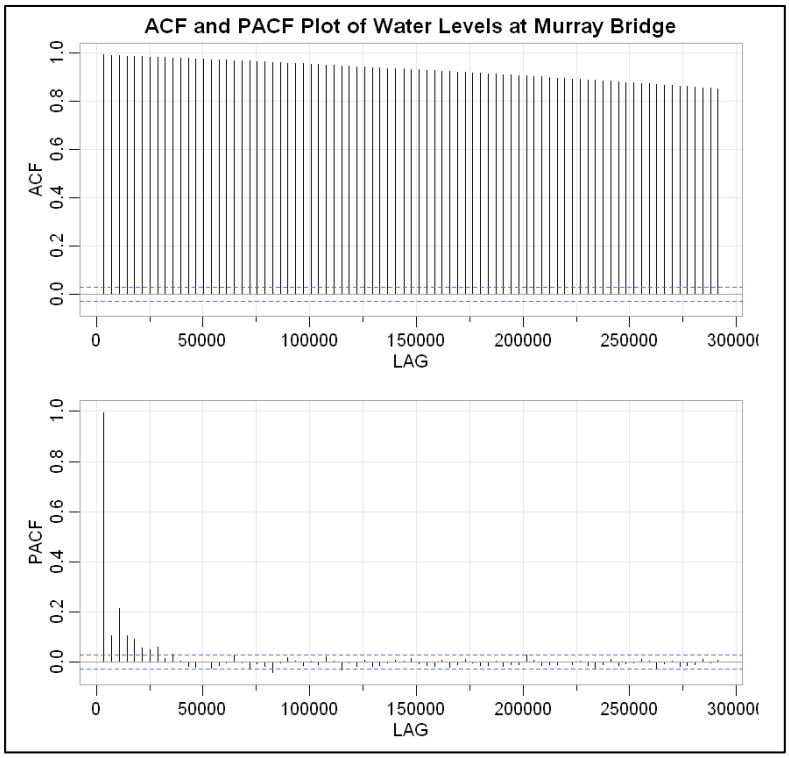

Figure 65: ACF and PCF plot of Water Level at Murray Bridge and the pattern of ACF indicates that there might be some seasonality in the data

ARIMA(0,d,q) model is selected when the PACF of the time series is sinusoidal or exponentially dampening, and there is a measurable spike at lag q in the plot of ACF, however none after the lag q (Rob, 2017). More ACF and PACF plots of different locations and parameters have been provided in Appendix-A.

3) Box-Ljung test:

Box-Ljung is a Portmanteau test where a more professional test for autocorrelation is conducted to check whether the residuals of a model have any autocorrelation or not (Rob & George, 2021e). The formula of the Box-Ljung is given in Figure 66.

$$Q^* = T(T+2)\sum_{k=1}^{\ell}(T-k)^{-1}r_k^2.$$

Figure 66: Formula of Box-Ljung Test where r_k is the autocorrelation of k times lagged time series

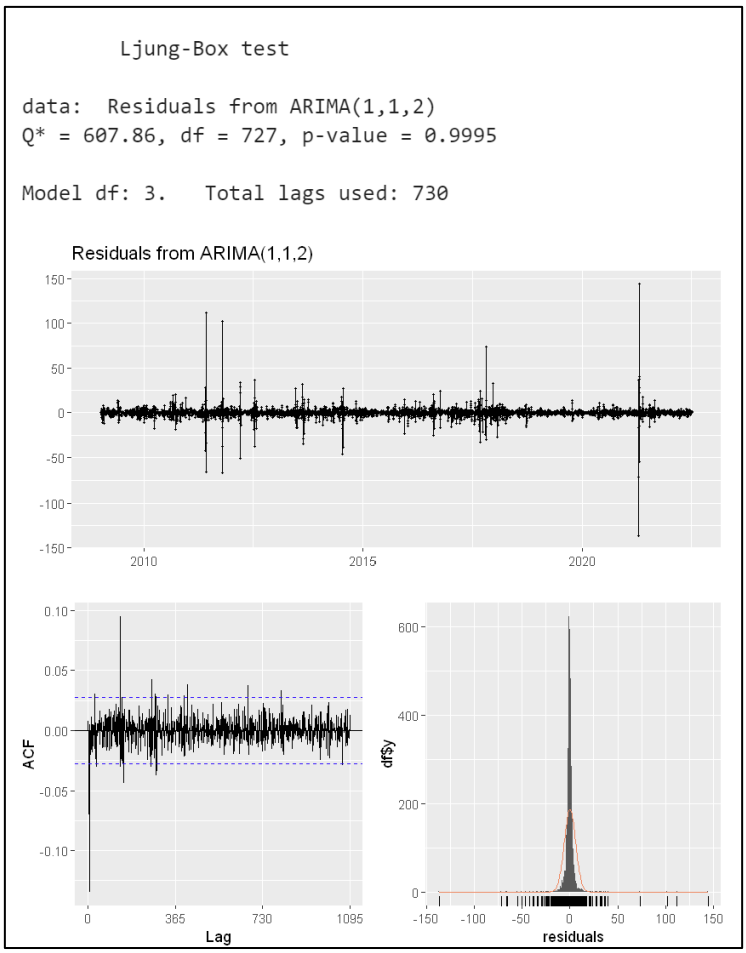

Figure 67: Ljung-Box test of Salinity Levels at Colignan informs us that residuals are normally distributed, and the p-value indicates that there is insignificant autocorrelation between time series values

Here, l is the largest lag considered for autocorrelations, r_k is the autocorrelation for lag k, and T is the number of observations. Furthermore, a large value of Q^* suggests that the autocorrelations are just a White Noise Time-Series. A Large value here means that the Q^* will have a χ^2 (chi-square) distribution (Rob & George, 2021e).

4) Akaike's Information Criterion (AIC):

The Akaike's Information Criterion is a measure of predictive accuracy to check which model might be the best (Rob & George, 2021d). The formula of AIC is given in Figure 68.

$$\text{AIC} = T\log\left(\frac{\text{SSE}}{T}\right) + 2(k+2),$$

Figure 68: Formula of AIC where k is the coefficient for the estimators

Here SSE is the minimum sum of squared errors, T is the number of observations for prediction, and k is the coefficients for the estimators. The main idea behind AIC is to chastise the fit of the model with the number of parameters that are required to be predicted. The model with the minimum value of AIC is often considered the best (Rob & George, 2021d).

5) Corrected Akaike's Information Criterion (AICc):

The Corrected Akaike's Information Criterion is utilized for small values of T because AIC selects a lot of estimators than required. Therefore, AICc is the bias-corrected version of AIC. Also, the model with a minimum value of AICc is the best model (Rob & George, 2021d). The Formula is given in Figure 69.

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2(k+2)(k+3)}{T-k-3}.$$

Figure 69: Formula of AICc where T is the number of observations for prediction

6) Schwarz's Bayesian Information Criterion (BIC):

Schwarz's Bayesian Information Criterion is similar to AIC. The only difference is that it chastises the fit of the model with the number of parameters that are required to be predicted more heavily. Moreover, the best model is the one with a minimum value of BIC (Rob & George, 2021d). The formula of BIC is given in Figure 70.

$$\mathrm{BIC} = T\log\left(\frac{\mathrm{SSE}}{T}\right) + (k+2)\log(T).$$

Figure 70: Formula of BIC where SSE is the minimum sum of squared errors

All the above parameters, specifically ACF, PCF, Box-Ljung Test, AIC, AICc and BIC, are implemented by the Acf, checkresiduals and glance methods of forecast Library in R Programming Language (Hyndman et. al, 2022).

Evaluation Metrics

There are several evaluation metrics for regression, such as Mean absolute error (MAE), Mean squared error (MSE), Root mean squared error (RMSE), Root mean squared logarithmic error (RMSLE), Mean percentage error (MPE), Mean absolute percentage error (MAPE) and R-square (R^2).

However, for this research, we would be only utilizing RMSE and MAPE because these are two commonly used scale-dependent measures. The scale-dependent measures are those in which the time-series data is on the same scale as the forecast errors. Furthermore, a forecast error (e_t) is the difference between an actual value and its predicted value (Rob and George, 2018f). The formula of RMSE and MAPE is given in Figure 71 and Figure 73, respectively.

$$\mathrm{RMSE} = \sqrt{\mathrm{mean}(e_t^2)}.$$

Figure 71: Formula of RMSE where e_t is the forecast error

$$p_t = 100e_t/y_t.$$

Figure 72: Formula of percentage error where e_t is the forecast error and y_t is the training data

$$\mathrm{MAPE} = \mathrm{mean}(|p_t|).$$

Figure 73: Formula of RMSE where p_t is the percentage error

APPENDIX – E

Support Vector Machine

To understand SVM for regression, let us consider a training set given in Figure 74 where R^n is the real coordinate space of n-dimension (Lin et al., 2006).

$$\{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\} \subset R^n \times R$$

Figure 74: Training Dataset where y is a response variable, and x is a predictor variable

The SVM model estimates the function displayed in Figure 75, where w and b are vector coefficients. Besides, <w,x> represents the dot product between w and x (Lin et al., 2006; Vapnik, 1998).

$$f(x) = < w, x > + b \quad w, x \in R^n, \quad b \in R$$

Figure 75: Function being estimated by the SVM algorithm where w and b are vector coefficients

$$\frac{1}{2}\|w\|^2 + C \cdot R_{emp}[f]$$

Figure 76: Regularized risk function formula where the second term is the empirical error

The function mentioned above is estimated by minimizing the regularized risk function given in Figure 76. The first part of the equation given in Figure 76 is known as the regularized term, whereas the second term is called the empirical error. Minimizing the regularized term will flatten a function and hence would act as a function capacity controller (Lin et al., 2006).

The empirical error term is measured with help of the loss function. Moreover, the C term is known as the regularization constant, and it decides the deviations from the loss function that would be tolerable (Lin et al., 2006). The loss function, which is ε-insensitive, is given in Figure 77.

$$L_t(y_i, f(x_i)) = \max\{0, |y - f(x)| - \varepsilon\}$$

Figure 77: Loss function formula where ± ε is the margin around the hyperplane

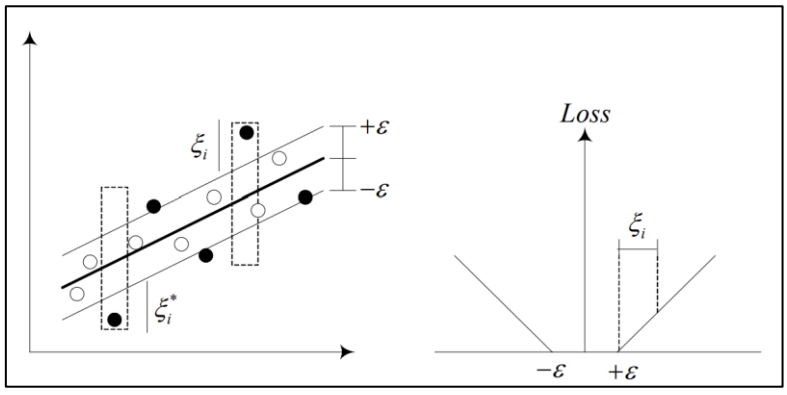

Figure 78: Loss setting for soft margin for loss function and linear SVM

The loss function defines a tube or margin of thickness 2*ε. Furthermore, the loss is zero when any predicted value lies in the ε-tube, and it is the difference between the radius ε of the margin and the predicted value when the predicted value is outside the ε-tube (Lin et al., 2006).

Let us assume that training data *X* is estimated with a precision of ε by a function *f*. In this particular case, it is believed that the problem is feasible. However, when the problem becomes infeasible, then slack variables

ξ and ξ* presented in Figure 78 are introduced to handle the infeasible restrictions of the optimization problem (Lin et al., 2006).

The formalized form of the above problem is given in Figure 79 and Figure 80.

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right)$$

Figure 79: Formalized version of the regularized risk function

$$\text{subject to} \begin{cases} y_i - < w, x_i > -b \le \varepsilon + \xi_i \\ < w, x_i > +b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

Figure 80: Conditions attached with the formalized risk function

The main objective now is to build a Lagrangian (also known as Lagrange function) from the formalized risk function mentioned in Figure 79 and its respective conditions presented in Figure 80 by adding a dual set of variables. Figure 81 displays that the Lagrange function with respect to the dual and primal variables has a minimax point at the solution (Lin et al., 2006).

$$L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{l}\left(\eta_i\xi_i + \eta_i^*\xi_i^*\right) - \sum_{i=1}^{l}\alpha_i\left(\varepsilon + \xi_i - y_i + < w, x_i > +b\right) - \sum_{i=1}^{l}\alpha_i^*\left(\varepsilon + \xi_i^* - y_i + < w, x_i > +b\right)$$

Figure 81: Lagrange function *L* with η, α as Lagrange multipliers

In Figure 81, *L* is the Lagrange function, and Lagrange multipliers are represented as η_i, $(\eta_i)^*$, α_i, and $(\alpha_i)^*$. The positive conditions that the dual variables in Figure 81 have to satisfy are given in Figure 82.

$$\eta_i, \eta_i^*, \alpha_i, \alpha_i^* \ge 0$$

Figure 82: Positive conditions that dual variables have to satisfy

To optimize the Lagrange multipliers (α_i, and $(\alpha_i)^*$), the above problem can be transformed into a dual problem. A quadratic objective function formed of α_i and $(\alpha_i)^*$ variables having one linear condition is contained in the dual problem, as shown in Figure 83.

$$\max W\left(\alpha_i, \alpha_i^*\right) = -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right) < x_i, x_j > -\varepsilon\sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right) + \sum_{i=1}^{l} y_i\left(\alpha_i - \alpha_i^*\right)$$

Figure 83: Quadratic objective function formed of α_i and $(\alpha_i)^*$ Lagrange multipliers

$$\sum_{i=0}^{l}\left(\alpha_i - \alpha_i^*\right) = 0 \text{ and } \alpha_i, \alpha_i^* \in [0, C]$$

Figure 84: The quadratic objective function is subjected to the above condition

After performing manipulations with the help of dual optimization and the Lagrange multiplier, we get:

$$w = \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)x_i$$

Figure 85: The solution of vector coefficient *w* from the equation given in Figure 28

Then put the value of the *w* vector coefficient provided by the equation in Figure 85 in the equation of Figure 76, we get:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) < x_i, x > + b$$

Figure 86: The solution function that is estimated by Linear SVM

Radial Basis Function

The SVM model explained above performs linear regression. However, when the need to optimize non-linear functions arises, the linear approach should be modified. This particular task is achieved by linearizing the relationship between y_i and x_i by replacing x_i with a mapping $\phi(x_i)$ into the feature space. The regression solution can be obtained by utilizing the original approach in feature space (Lin et al., 2006). Therefore, the equation in Figure 86 transforms into the equation provided in Figure 87 when the mapping function is used.

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(x_i, x) + b \quad \text{with} \quad K(x_i, x) = < \varphi(x_i), \varphi(x) >$$

Figure 87: Linear SVM transformed into Non-linear with the help of kernel function *K*

The mapping can be simplified with the aid of Kernel function *K* in Figure 87 by mapping the data inherently into the feature space without having complete knowledge about ϕ, which makes the whole process efficient (Lin et al., 2006). The commonly used kernels are provided in Table 45 below.

Table 45: Commonly utilized kernel functions

Kernels	Functions	Parameters
Linear	$<x,x_i>$	
Polynomial	$(<x,x_i> + 1)^d$	d
Radial basis function	$\exp\left(-\|x - x_i\|^2 / 2\sigma^2\right)$	σ
Sigmoid	$\tanh(b <x,x_i> + c)$	b, c

The approximation errors are equal to or greater than ε for data points linked with the kernel functions. These data points are called support vectors. Furthermore, the value of ε is inversely proportional to the number of support vectors (SV). Therefore, if the value of ε increases, the number of SV decreases, causing the representation of the solution to be less dense. However, the approximation accuracy placed on training data points plummets with the surging value of ε. In this case, ε is a compromise between closeness to the data and the paucity of representation (Lin et al., 2006; Cao & Tay, 2003).

The value of ε (margin) and C (cost or regularization constant) have to be tuned by the user, unlike the Lagrange multipliers, which are tuned by the optimization problem (Lin et al., 2006).

For this research, we would be utilizing the Radial basis function (RBF) as the kernel function because the data points are mapped by RBF into higher dimensional space nonlinearly, unlike the linear kernel function, which is a special case of RBF. More notably, RBF can take care of situations where the relationship between attributes and class labels is non-linear (Lin et al., 2006). Keerthi & Lin (2001) observed in their study that the performance of the linear kernel with cost parameter *C* was similar to that of the RBF kernel for certain values of C and σ. Similarly, the sigmoid kernel function operates like RBF for certain values of parameters (Lin & Lin, 2003). The RBF kernel has fewer hyperparameters than the Polynomial kernel, due to which the RBF kernel requires fewer numerical computations than polynomial kernels, whose values might go to zero or infinity (Lin et al., 2006). Hence, for this research RBF kernel is chosen.

The SVM with RBF kernel is implemented by the svm_rbf function present in the parsnip library of the R programming language (Kuhn, 2022a).