Effect of Global Warming on Long-Term Sea Level Change in Tuvalu

ESE527: Practicum in Data Analytics & Statistics

Team H-H: Huanyue Liao and Hanyu Wang
Washington University in St. Louis
McKelvey School of Engineering

05/07/2023

Content Table

Executive Summary

Motivation

Tuvalu is a two-seasoned island country located in the western central Pacific Ocean and has been known for its low-lying nature, with its lowest point being at the sea level and the highest point rising only 4.6 meters above the sea level (Aung *et al.* 2009). As global warming becomes a pressing issue in recent times, the rise in the Earth's surface temperature has resulted in the melting of glaciers and ice sheets, leading to an expansion of ocean water and a rise in sea levels, while all these changes are projected to cause the sinking of low-lying areas such as Tuvalu (Riebeek 2010; *Current and Future Climate of Tuvalu* 2014). According to Cooley *et al.* (2022), the global mean sea level has increased by 0.2 m since 1901 and is expected to continue rising, resulting in a higher frequency of extreme sea levels. This project is then proposed to estimate the trend of the elevation of sea level to predict the sinking of Tuvalu.

The change in sea level is a result of the complex combination of a series of influences, including daily tides, meteorological effects, thermal effects, seismic activity, oceanographic effects, and vertical land movement (Aung *et al.* 2009). Among all, El Niño-Southern Oscillation (ENSO) events, which belongs to the oceanographic effects, have the most significant short-term influence on Tuvalu sea level (Singh and Aung 2005), while global warming (the thermal effects) is believed to be the most crucial factor influencing the long-term sea level change in Tuvalu (*Current and Future Climate of Tuvalu* 2014). For that reason, this project will focus on the effect of temperature change due to global warming on the sea level of Tuvalu instead of the many other influences.

Decisions to be impacted

This project provides projections of future sea level rise due to global warming and its potential impacts on many decisions. Since the elevation of the sea level in Tuvalu is unpreventable, the nation will disappear one day and its residents will have to leave as climate refugees before their homes sink. This project will help the Tuvalu government, the international community, and other countries get prepared for the relocation and resettlement of the people by informing them of the time when the country will go underwater.

Besides, this project addresses the severity and the harmfulness of global warming, alerting people that the impact of global warming is far greater than merely the increase in temperature (Riebeek 2010). Worthwhile mentioning, ocean acidification is another problem caused by global warming since one-quarter of the carbon dioxide produced by human activities is absorbed by the oceans and increases the acidity of the ocean, which significantly impacts the marine ecosystem, including the health and survival of coral reefs while they are also vulnerable to heat stress (*Current and Future Climate of Tuvalu* 2014). Coral reefs are essential to Tuvalu's economy since they provide critical habitats for fish and protect the country from storms and erosion, and also essential to the existence of Tuvalu since Tuvalu is made up of nine atolls that are formed from coral reefs. When coral reefs die, it will result in the loss of fish stocks, which are a crucial source of protein for Tuvalu communities, and about 67% of families are involved in fisheries (Gallagher 2019). By predicting the trend of sea level rise, this project aims to raise people's awareness of the vulnerabilities of Tuvalu to rising sea levels due to climate change. This project also advocates for normal people as well as the international community to take action to mitigate greenhouse gas emissions to slow down global warming.

What's more, predicting the sea level of Tuvalu provides the government with information on the actions they need to take. The government may build sea walls and elevate the floor of the buildings in response to the rise of sea level. This project makes the timeline of the adaptations available to the government. At last, Tuvalu is not the only island country on the verge of going underwater. The rising sea level is a threat to other Pacific Ocean countries and Indian Ocean countries such as Maldives. This study can serve as an example to provide the method of predicting the sea level change and forecasting the sinking.

Business value

The business value of this project is intertwined with its impacts described above. Firstly, based on the disappearance of Tuvalu, some supply chains that relied on this nation such as fisheries will have to be adjusted, while this project provides a time frame for the adjustment (*Fisheries Department* Accessed 2023). As a side note, the value of fisheries exports is about three times larger than fisheries imports in Tuvalu (NOAA Fisheries 2010). Secondly, when the residents in Tuvalu resettle into other countries, the markets and societies of these countries will be impacted and this study prepares the markets for the impact. Next, raising the harms of global warming can lead to innovations in the industry of renewable energy. Then, the adaptation that the government takes will create jobs for construction companies. Besides, addressing the problem that Tuvalu faces might attract global support including financial funding. Lastly, despite that Tuvalu is a small island country with limited economic value, there are other countries like Maldives that have the same crisis. Referring to this project could yield significant economic benefits.

Data Description and Preprocessing

Dataset overview

The prediction of the sinking of Tuvalu is made based on the historical sea level data and temperature data provided by the Australian Bureau of Meteorology (BOM). The BOM has installed tide gauges in Funafuti, the capital of Tuvalu, to measure the relevant data from March 1993 to February 2023.

Three time-series datasets are retrieved from the tide gauges from BOM. Specifically, they are the monthly sea level data, monthly air temperature data, and monthly water temperature data, each of which has a length of 360 rows. The data are initially recorded hourly and then processed into monthly data on the official website. The cleaned monthly sea level data has 357 rows, while the cleaned monthly air temperature data has 309 rows, and the cleaned monthly water temperature has 325 rows. After further eliminating the data affected by severe El Niño events, the monthly sea level data has 333 rows, the air temperature has 285 rows, and the water temperature has 301 rows. The three time-series datasets are composed of two different types of data, which are the desired values (sea levels and temperatures) and the dates on which the values are recorded. The values are numerical continuous variables, and the dates are numerical discrete variables.

The raw monthly sea level data, air temperature data, and water temperature data from BOM have been visualized using Python in Figure 1.

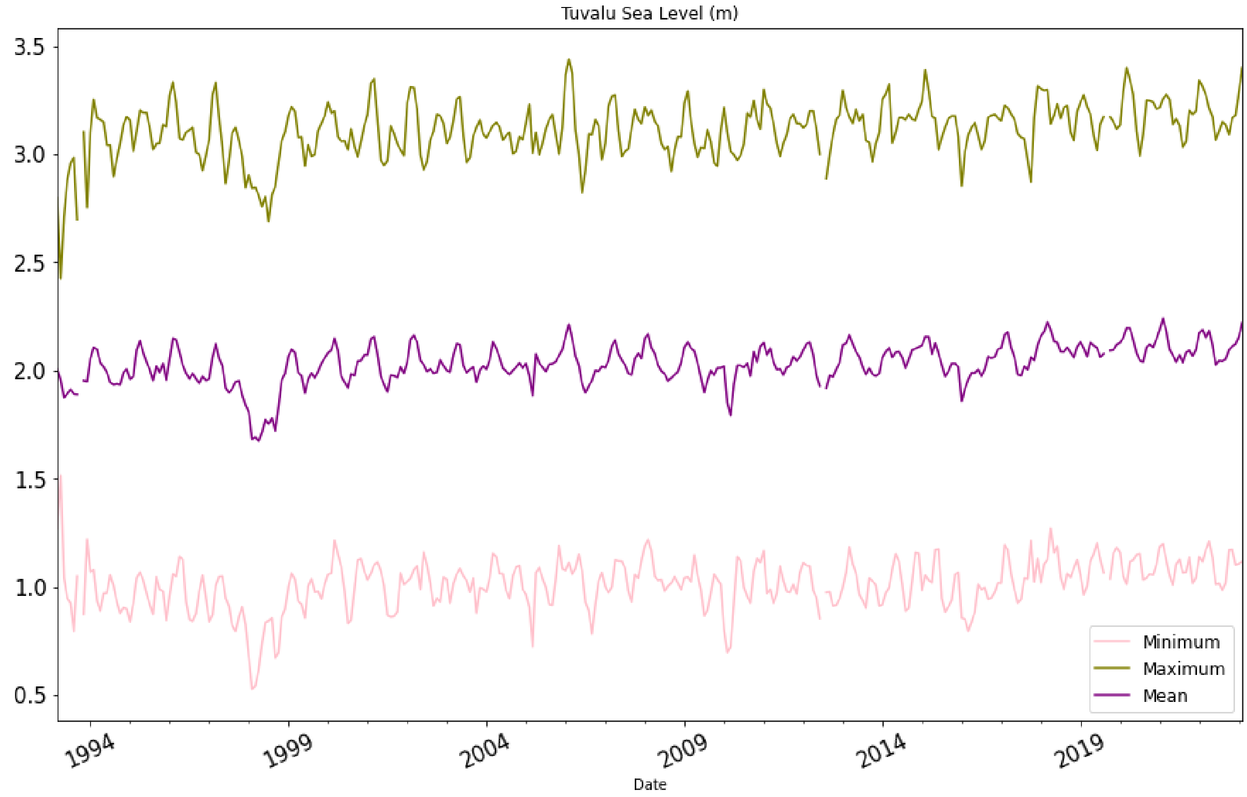


Figure 1a. The raw monthly Tuvalu sea level plot in meters.

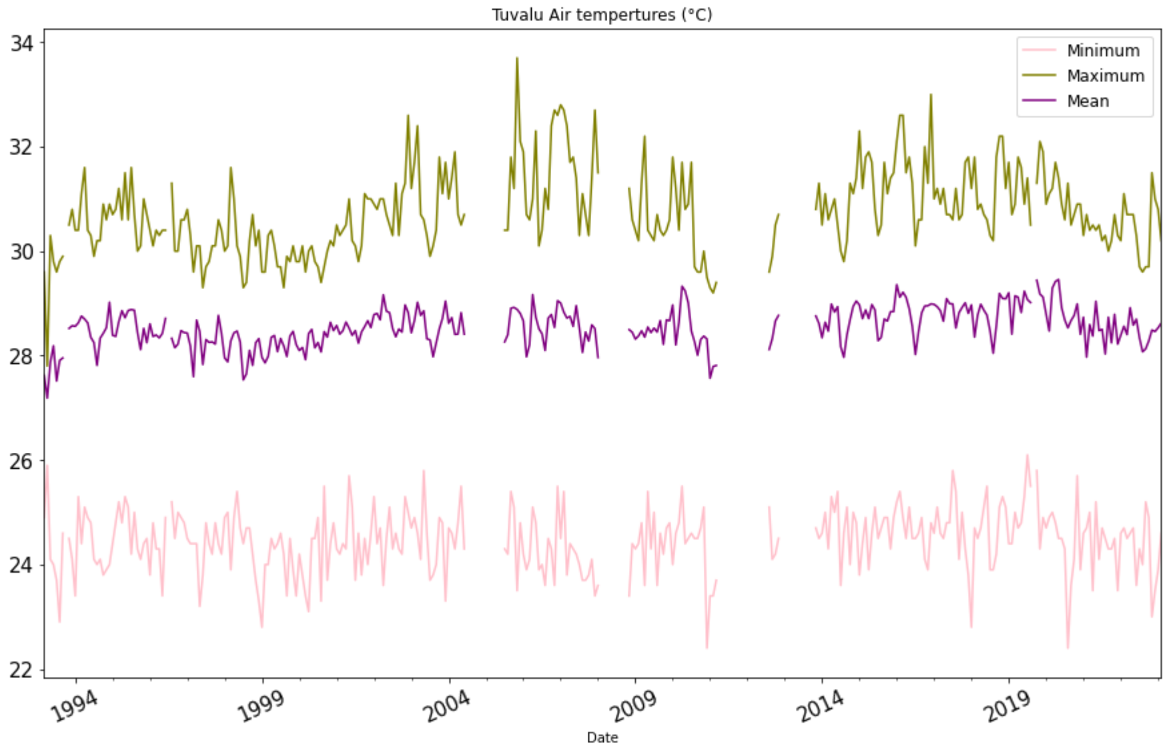


Figure 1b. The raw monthly Tuvalu air temperature in Celsius degrees.

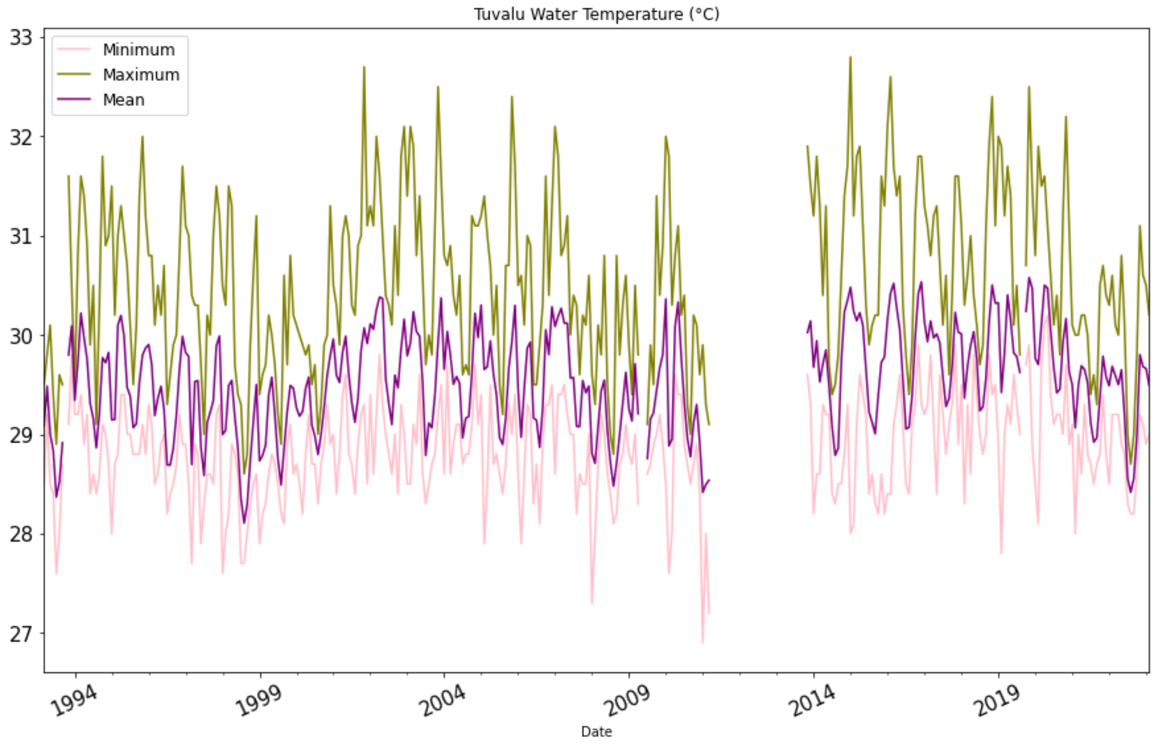
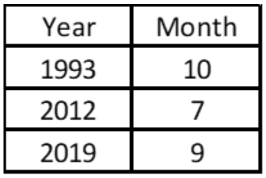


Figure 1c. The raw monthly Tuvalu water temperature in Celsius degrees.

The missing parts in Figure 1 represent the existence of invalid null data. Specifically, the missing data of the monthly sea level, the monthly air temperature, and the monthly water temperature are found and listed in Table 1.

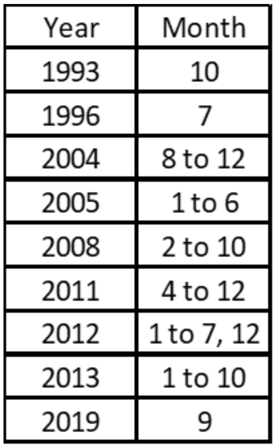
Table 1. **(a)** The null data of monthly sea level **(b)** The null data of monthly air temperature. **(c)** The null data of monthly water temperature.

(a)



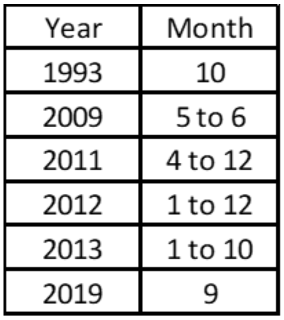
Year	Month
1993	10
2012	7
2019	9

(b)



Year	Month
1993	10
1996	7
2004	8 to 12
2005	1 to 6
2008	2 to 10
2011	4 to 12
2012	1 to 7, 12
2013	1 to 10
2019	9

(c)



Year	Month
1993	10
2009	5 to 6
2011	4 to 12
2012	1 to 12
2013	1 to 10
2019	9

Data cleaning

In light of the fact that time series data are applied for analysis in this project, STL decomposition and ARIMAX are used in the next section, which are methods robust to outliers. Hence, the data preprocessing period only eliminates some significant outliers influenced by El Niño events as indicated by the abnormal drop in Figure 1a.

As many studies suggested, there were many weak El Niño events that occurred during the last 30 years, and there was a severe one that occurred in March 1997 and lasted throughout 1997 and 1998 which caused the sea level in Tuvalu to drop significantly (Aung *et al.* 2009; Singh and Aung 2005; Hunter 2002). El Niño refers to unusual warming in the sea surface temperature and leads to positive barometric pressure and is often followed by La Niña, which refers to the cooling in sea surface temperature and negative barometric pressure (Singh and Aung 2005). The El Niño event in 1997 was very severe and the following La Niña was weak, so, the barometric pressure was abnormally positive during 1997 and 1998. The strong barometric pressure effect then became the most influential factor in changes to Tuvalu's sea level in 1997 and 1998 which explains the significant drop in sea level (Singh and Aung 2005). Since the topic of this project is about the effect of the increasing temperature due to global warming instead of the barometric pressure due to the El Niño event, the sea level data from March 1997 to December 1998 were considered outliers and were excluded in the following analysis.

Therefore, the data cleaning process in the report is done by eliminating the null data and rejecting the data affected by the severe El Niño event before 1998. Besides, due to the decisive nature of mean values, the minimum and maximum values are excluded for further analysis.

The cleaned data are visualized using R in Figure 2.

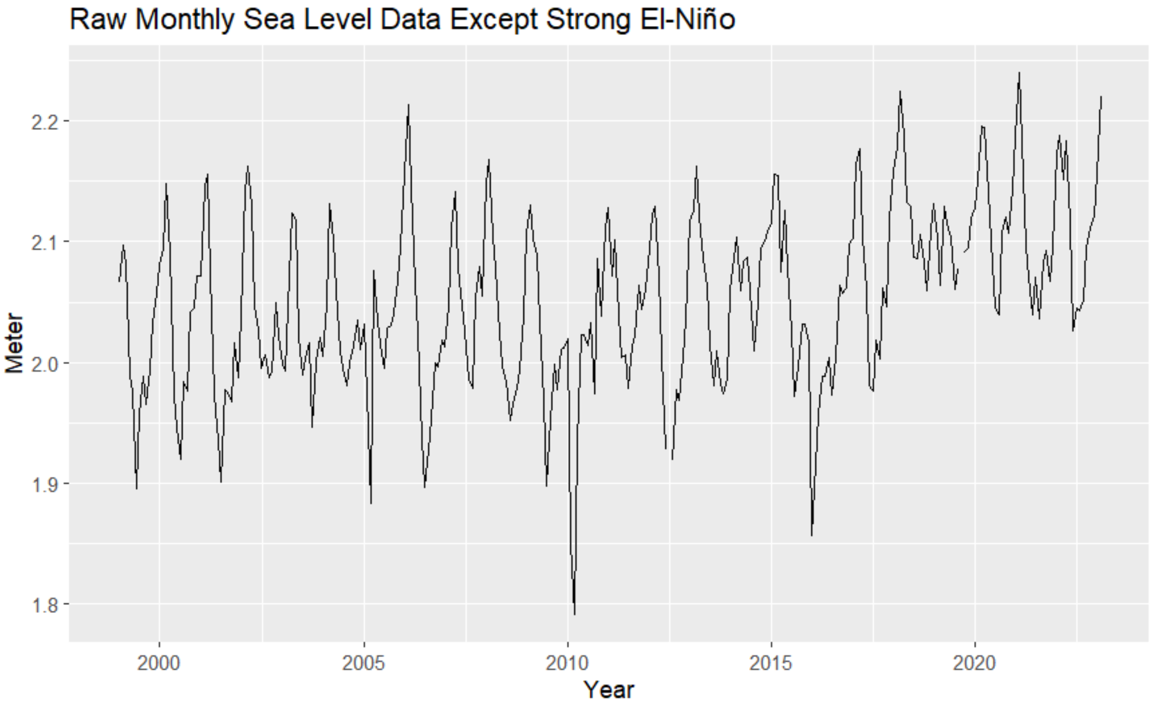


Figure 2a. The raw monthly Tuvalu sea level in meters excluding strong El Niño.

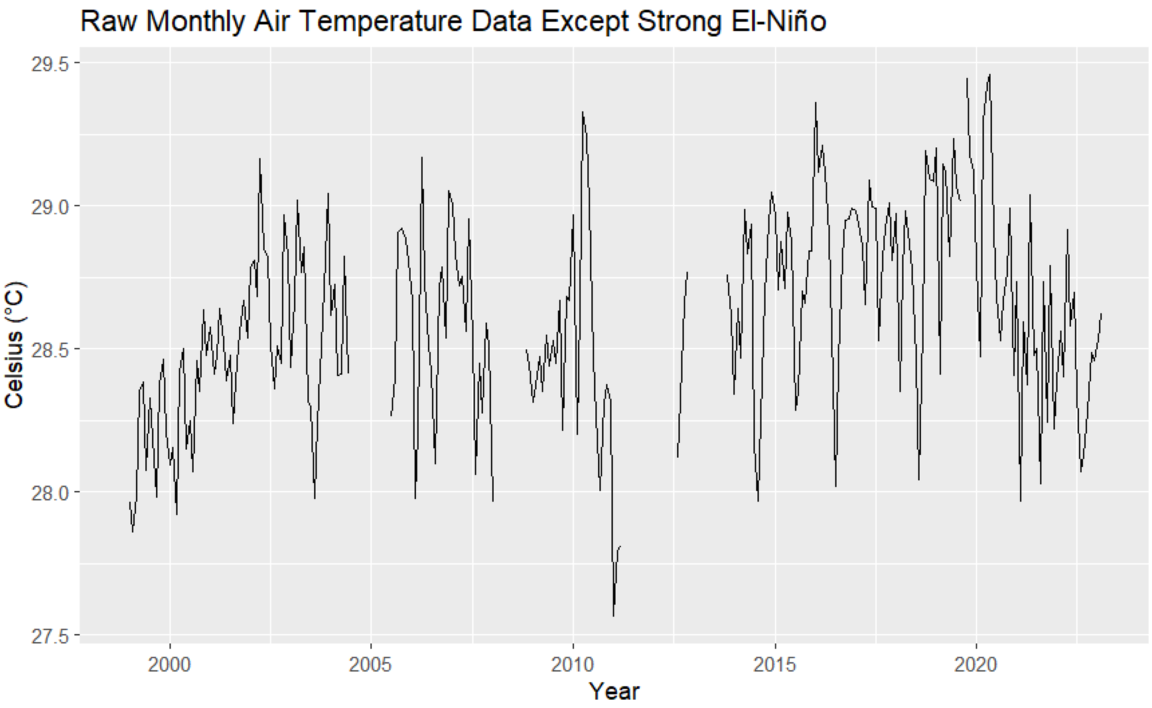


Figure 2b. The raw monthly Tuvalu air temperature in Celsius degrees excluding strong El Niño.

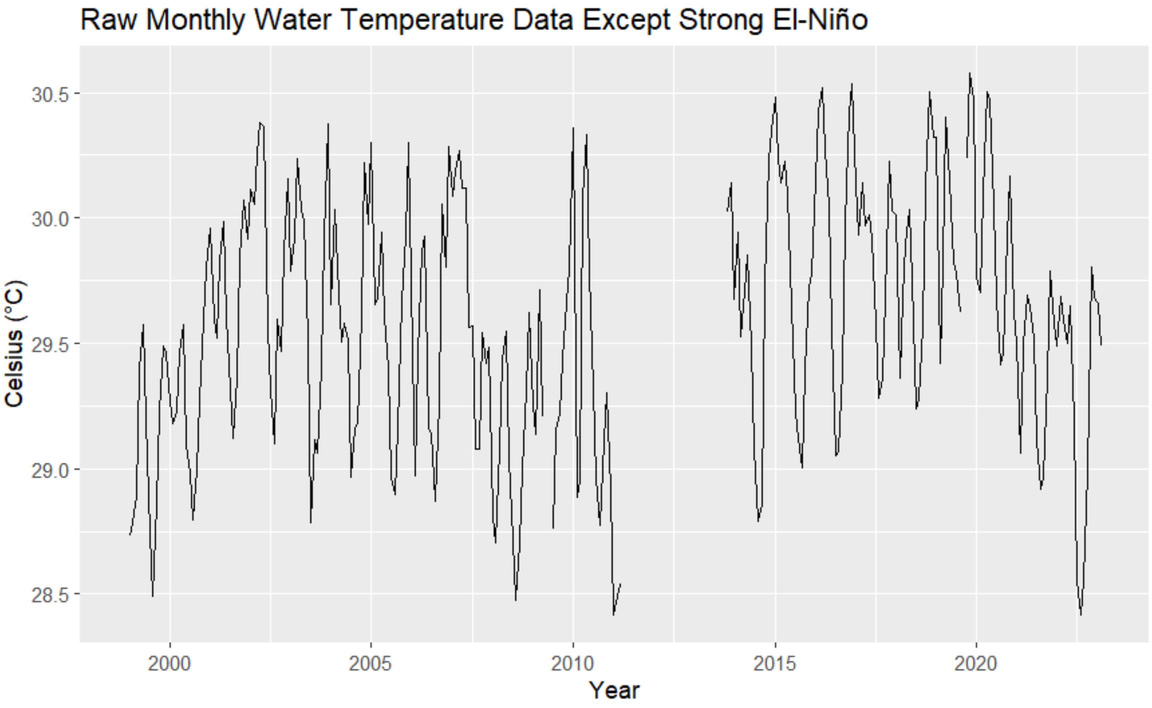


Figure 2c. The raw monthly Tuvalu water temperature in Celsius degrees excluding strong El Niño.

Since STL decomposition requires consecutive time series data, a built-in function *na_kalman* in the '*imputeTS*' package is applied using R. The function *na_kalman* uses Kalman smoothing and state space model to replace the null data. This approach leverages the temporal dependencies in the dataset to generate more accurate imputations compared to simpler methods such as linear interpolation. The Kalman filter algorithm estimates the hidden state of the time series, while the smoother algorithm improves the estimates by incorporating future observations.

Stationarity

The next step of data preprocessing is validating the stationarity of the data using the Augmented Dickey-Fuller (ADF) test. The ADF test is a widely used statistical test for determining whether a time series has a unit root, which is a characteristic of non-stationary time series. The test is based on a null hypothesis that the series has a unit root, and a rejection of this hypothesis suggests that the series is stationary. In the ADF test, if the t-statistic return is smaller than the given critical value (both are negative), then the null hypothesis can be rejected in the presence of a unit root in the data (Mushtaq 2011). In another way, if the p-value of the ADF test is less than the significance level ($\alpha = 0.05$), the null hypothesis can also be rejected (G 2023). In these scenarios, the data series does not have a unit root, and the dataset satisfies the stationarity criteria. The p-values of the data used in this report are all less than 0.01, so the datasets are stationary.

Train-test split

In order to further evaluate the performance of the models, the datasets are divided into training data and testing data at an 8:2 ratio. Specifically, the training dataset is used to perform the analysis and determine the parameters of the model, and the testing dataset is used to evaluate the forecasting result of the fitted model. The training data has a size of 232 months from January 1999 to April 2018, while the testing data has a size of 58 months from May 2018 to February 2023.

Modeling Approach

STL decomposition + Random walk

Seasonal-Trend decomposition using LOESS (STL decomposition) is a time series decomposition method that separates a time series into three components, which are trend, seasonality, and residual (also called as noise). It is a widely used method for time series analysis and forecasting. The trend component of a time series represents the long-term behavior or direction of the series. The seasonal component captures the regular, recurring patterns or fluctuations in the data within a fixed period, such as the yearly cycle in this project. The residual component is the part of the time series that cannot be explained by the trend and seasonal components, and it represents the random or irregular fluctuations in the data (Hyndman 2018).

The STL decomposition method is based on a regression technique called locally weighted scatterplot smoothing (LOESS) that fits a smooth curve through the points in sequence to estimate the trend and seasonal components. The method works by dividing the time series into smaller segments and applying the LOESS smoothing technique to each segment. This helps to capture local variations in the data while still preserving the overall trend and seasonality.

Random walk is a stochastic process formed by adding up independent, identically distributed random variables (Lawler and Limic 2010). That being said, there is no clear pattern in a random walk, while the future values are dependent on the past values. When the random walk and the STL decomposition model are combined, the remainder of the STL decomposition is modeled as a random walk. The resulting model is then a good fit to the data that is able to identify the unexpected spikes and dips as remainders.

Therefore, the STL decomposition + Random walk model is used in this project as descriptive data to analyze the three datasets (sea level, water temperature, and air temperature) to understand what has happened. The STL decomposition model is selected because the datasets used in this project are seasonal time-series data, while the random walk model is selected because of the existence of random values in the residual part. In Figure 3, the values of the decomposition components for each dataset are shown.

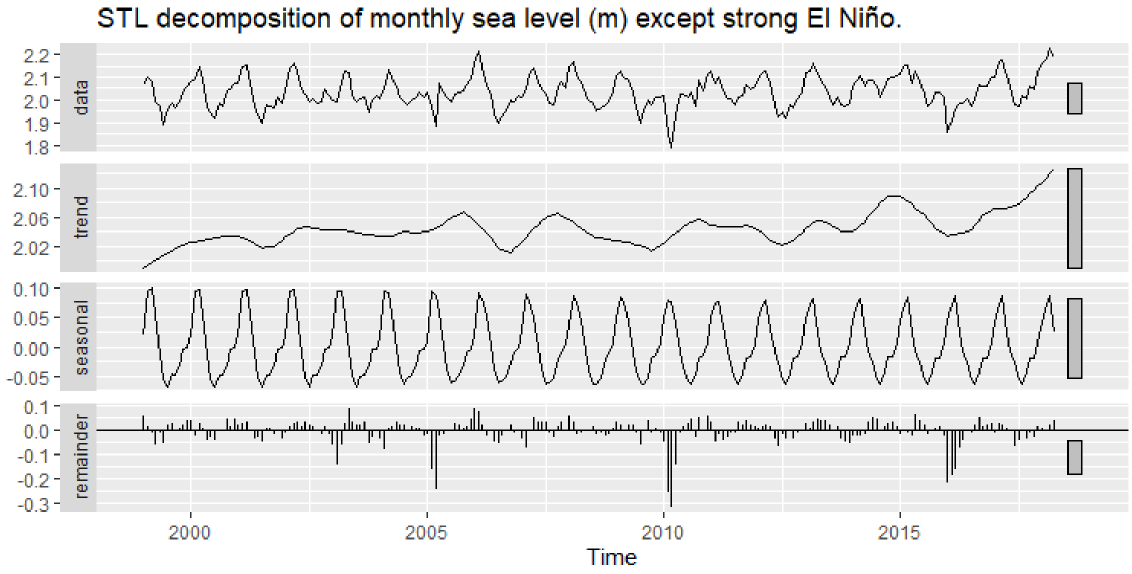


Figure 3a. The STL decomposition of Tuvalu sea level in meters excluding strong El Niño.

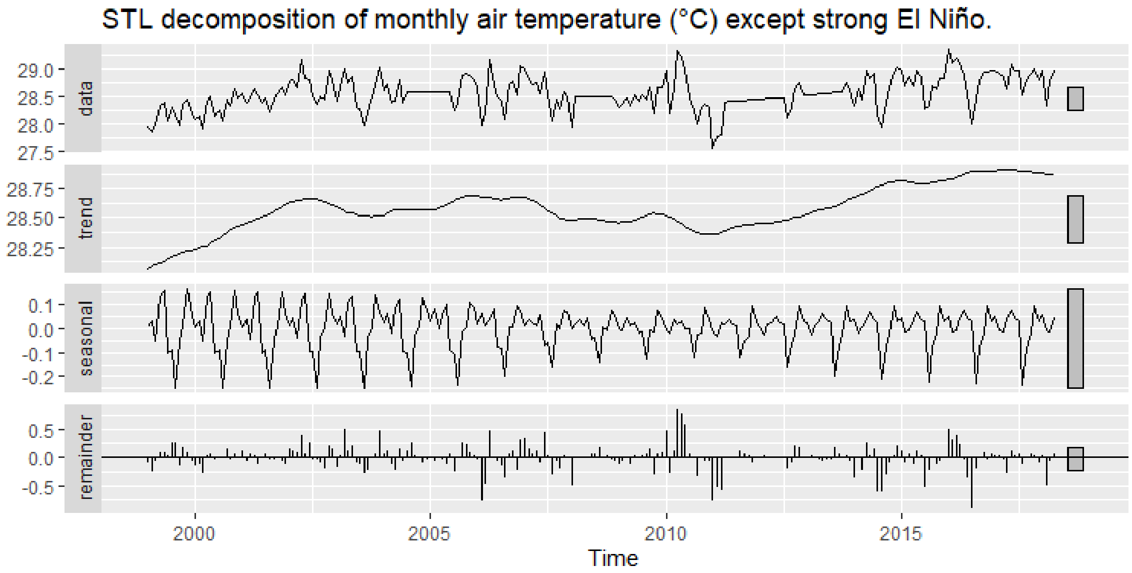

STL decomposition of monthly air temperature (°C) except strong El Niño.

Figure 3b. The STL decomposition of monthly Tuvalu air temperature in Celsius degrees excluding strong El Niño.

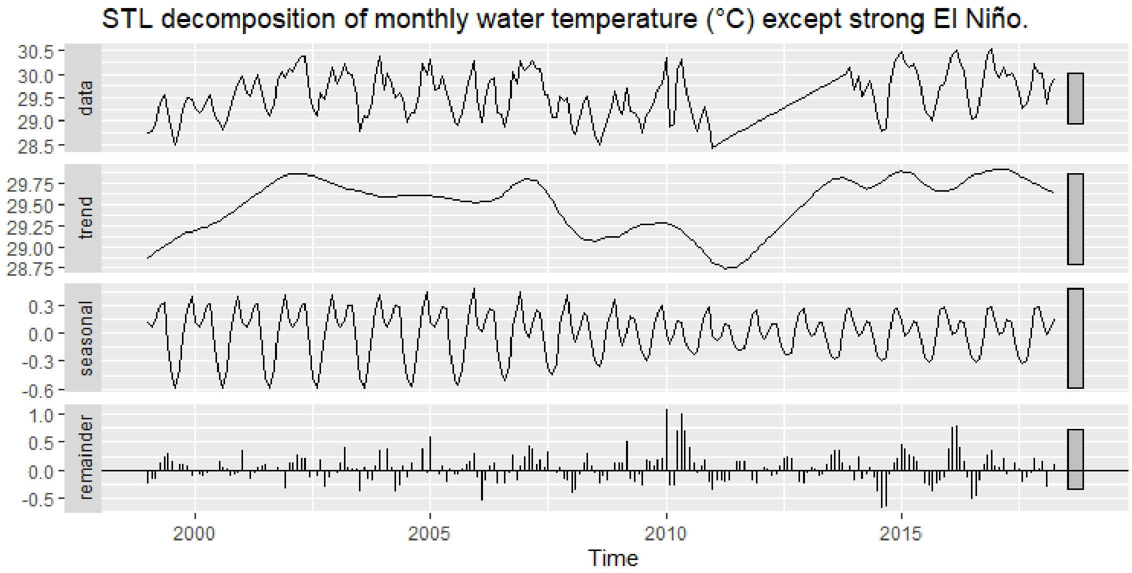

STL decomposition of monthly water temperature (°C) except strong El Niño.

Figure 3c. The STL decomposition of monthly Tuvalu air temperature in Celsius degrees excluding strong El Niño.

STL: Remainder

For further analysis, it is necessary to identify if the remainder requires additional decomposition. The first step is identifying the stationarity of the remainder. An ADF test is applied and the result indicates the remainders are stationary. Based on the result, an ARMA(p, q) model is applied to fit the residual instead of a MA(q) model.

The ARMA model refers to the Auto-Regressive Moving Average model while the AR model refers to the Auto-Regressive model and the MA model refers to the Moving Average model. As indicated in their names, the MA model is suitable for stationary time series data with a moving average pattern while the ARMA model is more flexible as it can handle both autoregressive and moving average patterns in the data. Therefore, when the remainder of the STL decomposition is stationary, an ARMA model is more suitable since it can capture the underlying patterns of the remainder in a better way.

Then, the function called *auto.arima* is used to optimize the ARMA model to minimize the Akaike information criterion (AIC). As a result, the remainders of the monthly sea level data and the water temperature data are fitted on the ARMA(2,1) model, where the $p = 2$ represents the order of the AR component and $q = 1$ represents the order of the MA component. At the same time, the remainder of the monthly air temperature data is fitted on the ARMA(1,0) model, with $p = 1$ and $q = 0$. The performance of the fitted ARMA models is visualized in Figure 4. The middle integers are 0 in the ARIMA models representing the models being ARMA models.

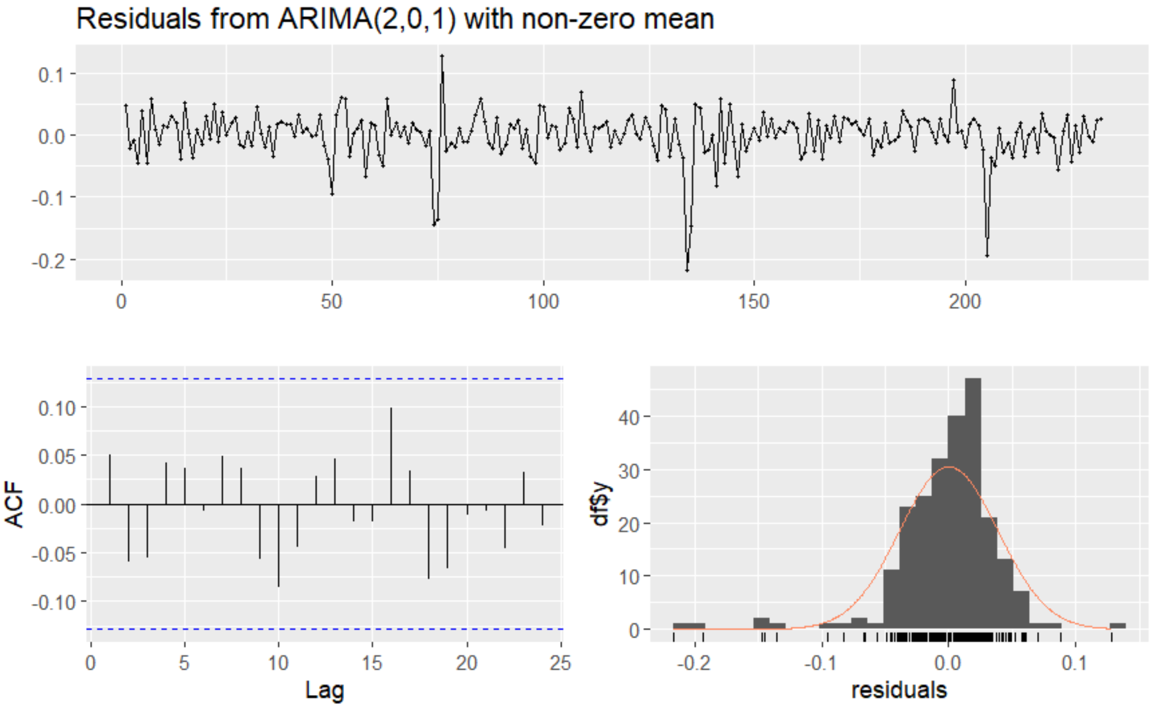


Figure 4a. The ARMA(2,1) model of the STL remainders of monthly Tuvalu sea level in meters.

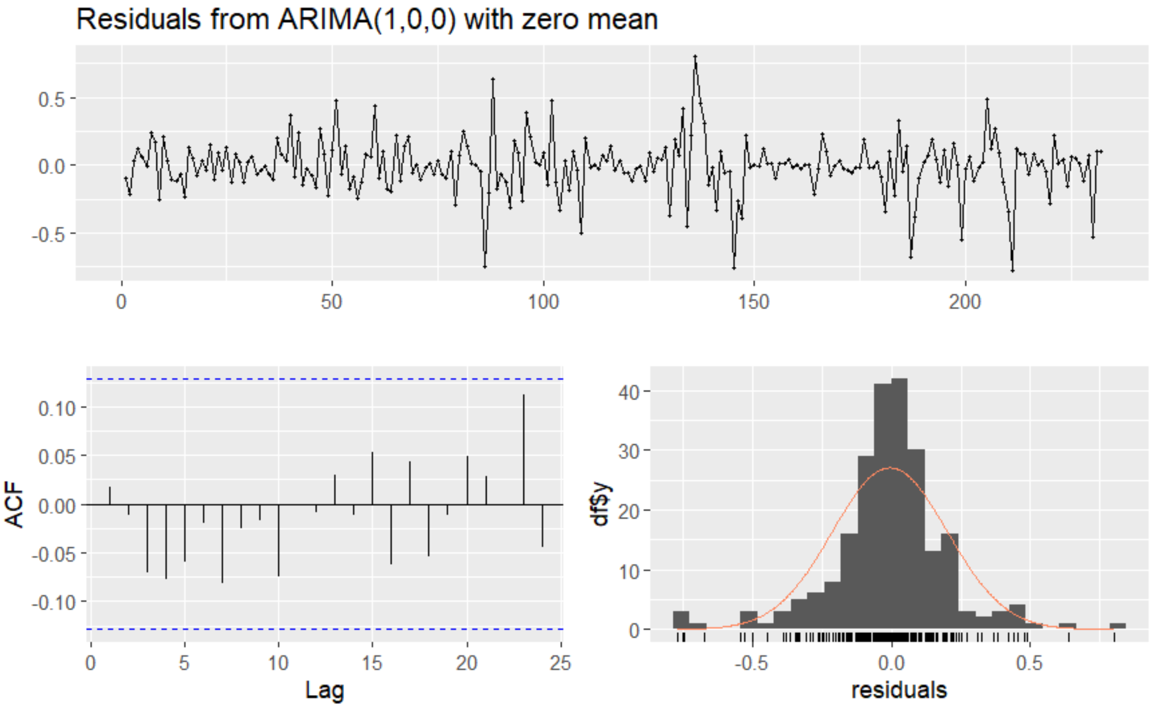


Figure 4b. The ARMA(1,0) model of the STL remainders of monthly Tuvalu air temperature in Celsius degrees.

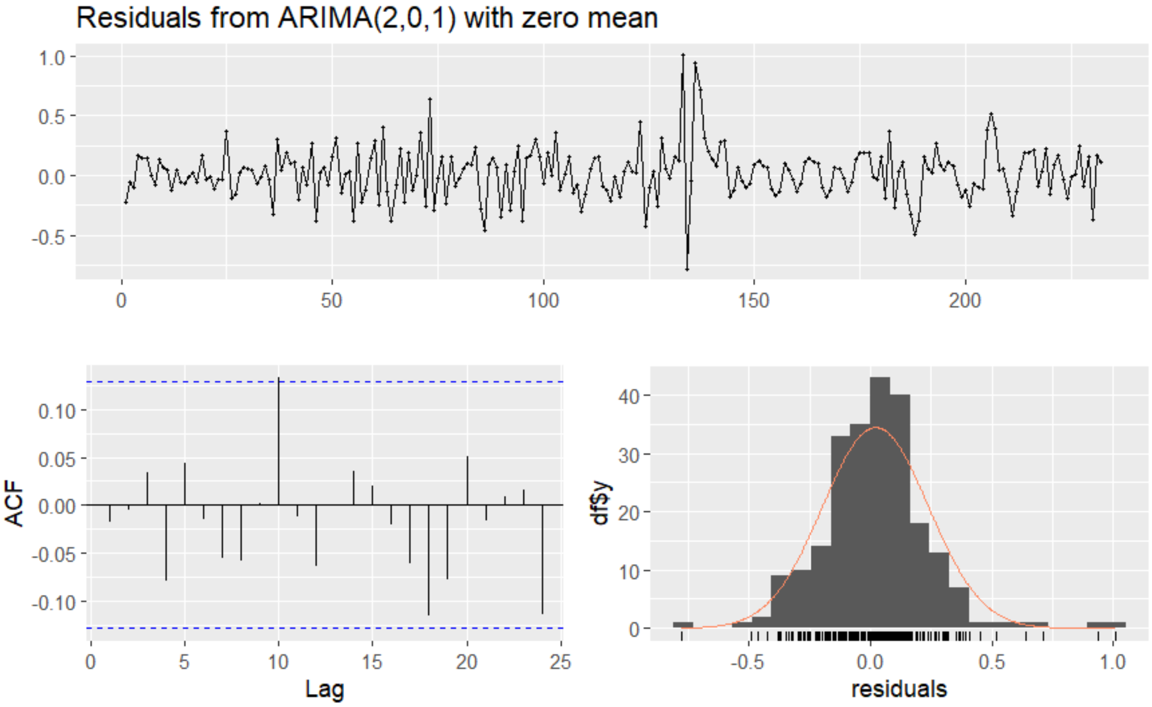


Figure 4c. The ARMA(2,1) model of the STL remainders of monthly Tuvalu air temperature in Celsius degrees.

From the above figures, it is evident the remainders all have a Gaussian distribution with their ACF being non-significant. To validate the results, a Ljung-Box test is applied to further check for autocorrelation of the residuals after ARMA model fits on the reminders. The null hypothesis of the Ljung-Box is that the residuals are independently distributed. As for the ARMA models' residuals of the monthly sea level reminders, water temperature reminders, and air temperature reminders, the p values of the Ljung-Box test are all larger than 0.05, indicating the null hypothesis cannot be rejected, the residuals of ARMA models are independently distributed. In this way, the residuals of ARMA models are simply white noises, and the ARMA models have captured all the underlying patterns, the ARMA model will be used to implement the reminder component from the STL decomposition.

STL: Seasonal

The seasonal component of the sea level, water temperature, and air temperature is shown below in Figure 5.

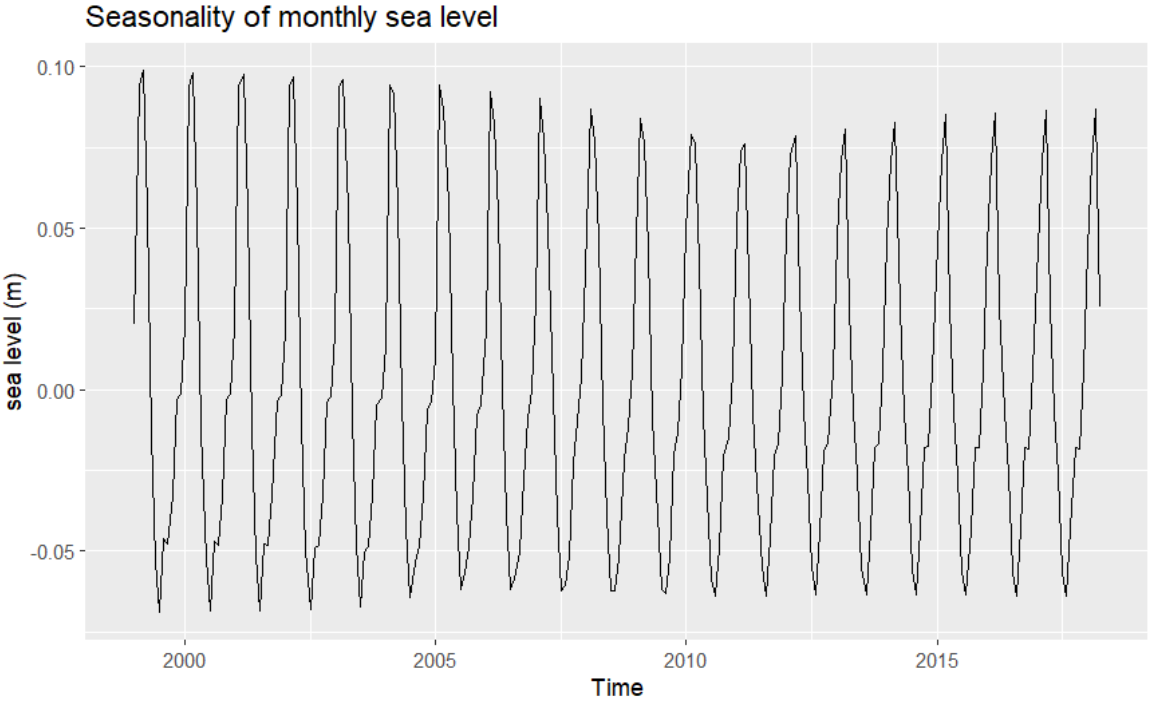


Figure 5a. The seasonal component of the monthly sea level data.

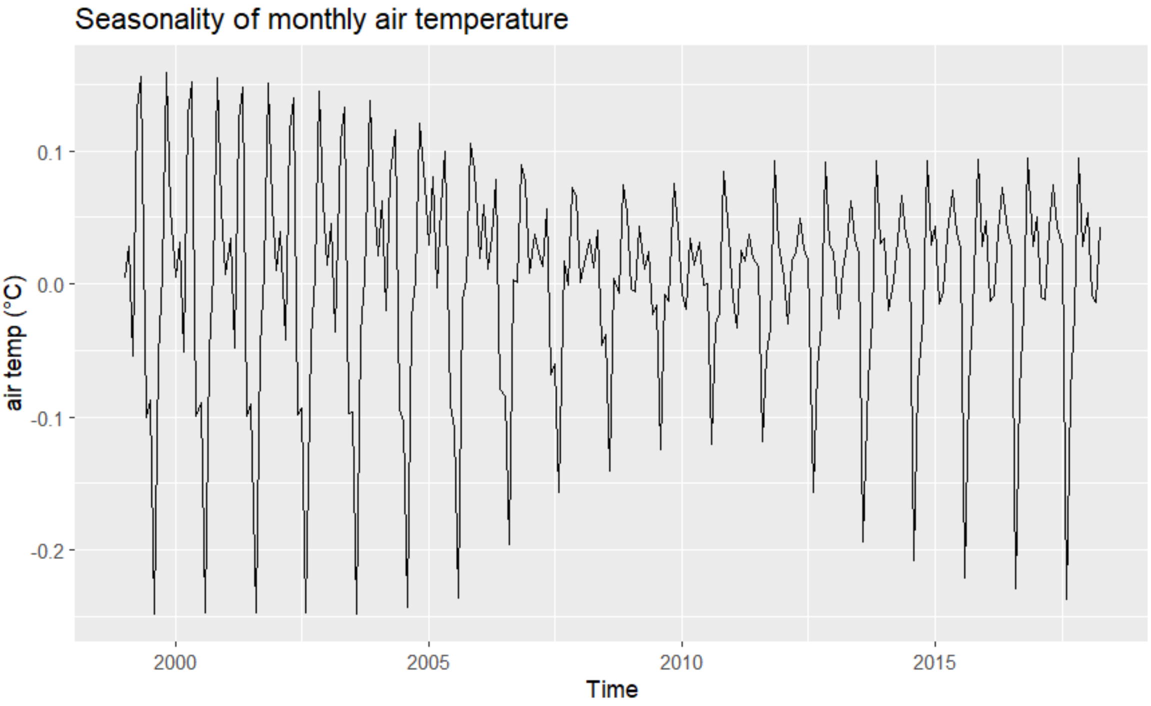


Figure 5b. The seasonal component of the monthly air temperature data.

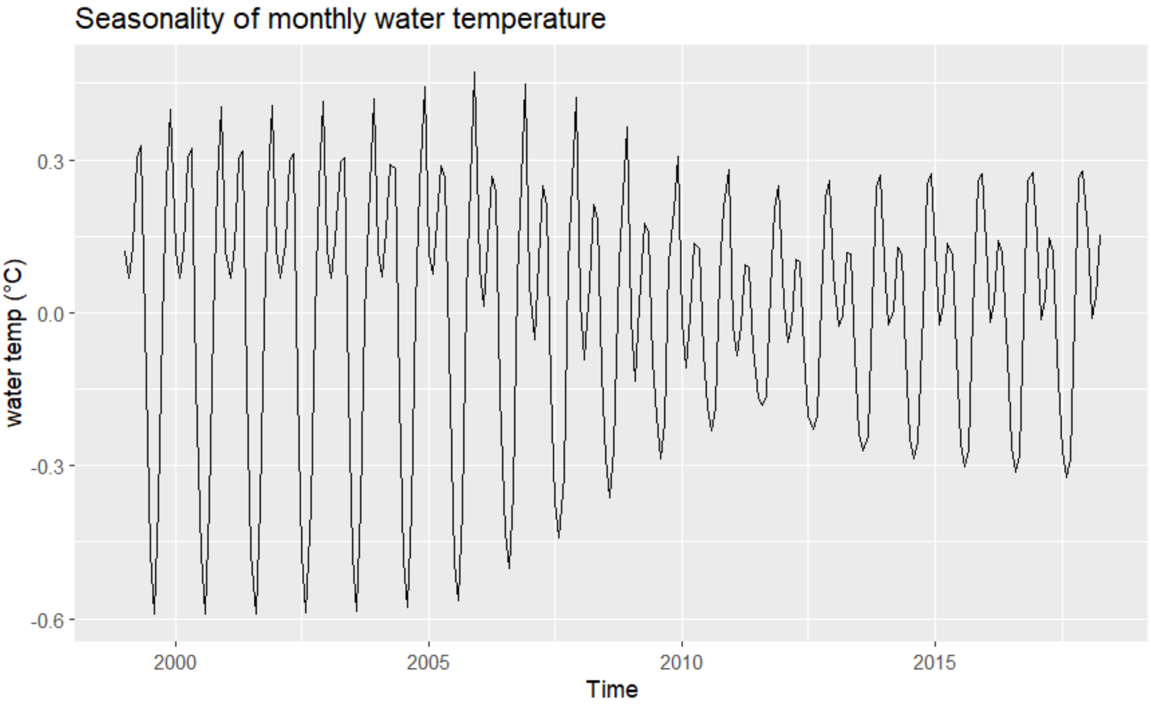


Figure 5c. The seasonal component of the monthly water temperature data.

To improve the clarity, the box plots of the data are presented in Figure 6.

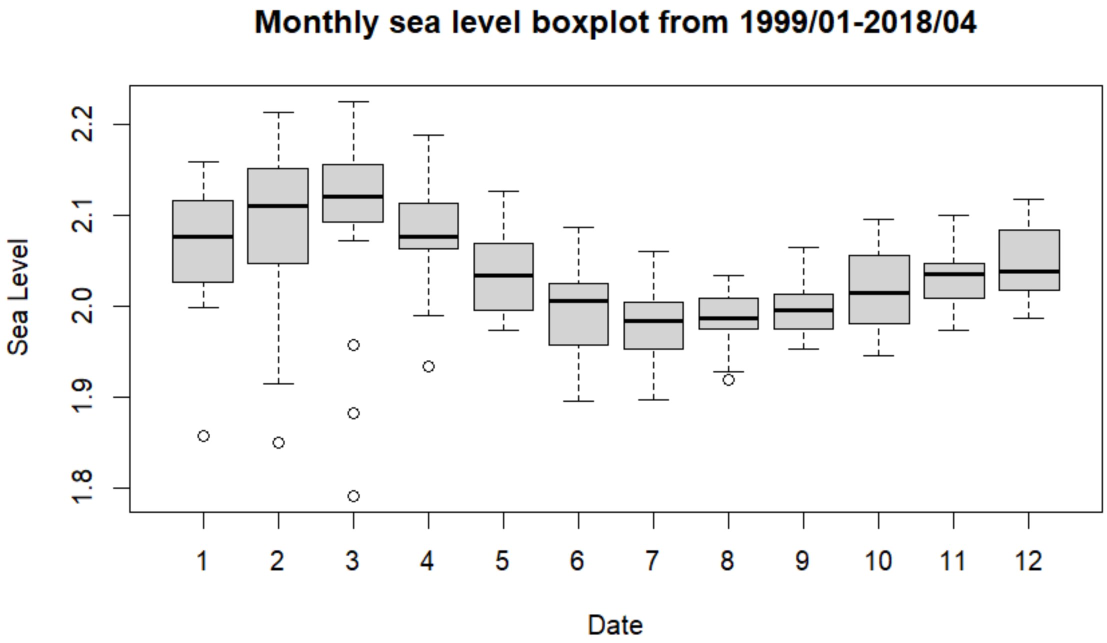


Figure 6a. The boxplot of monthly sea level within a cycle.

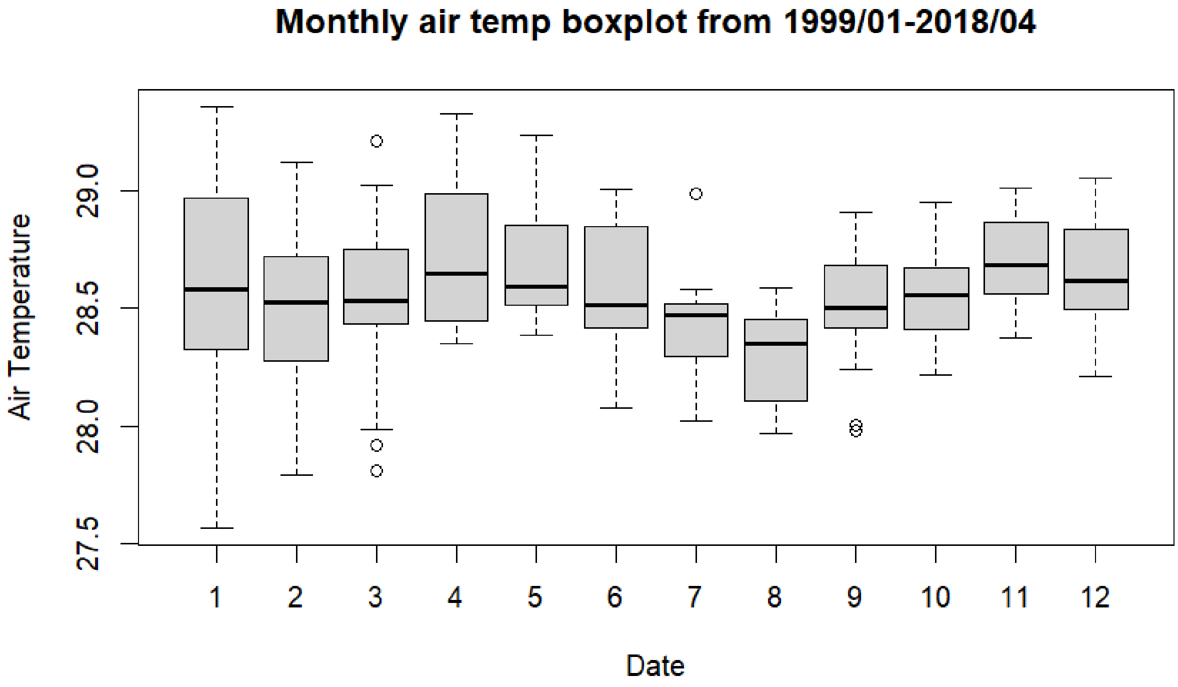


Figure 6b. The boxplot of monthly air temperature within a cycle.

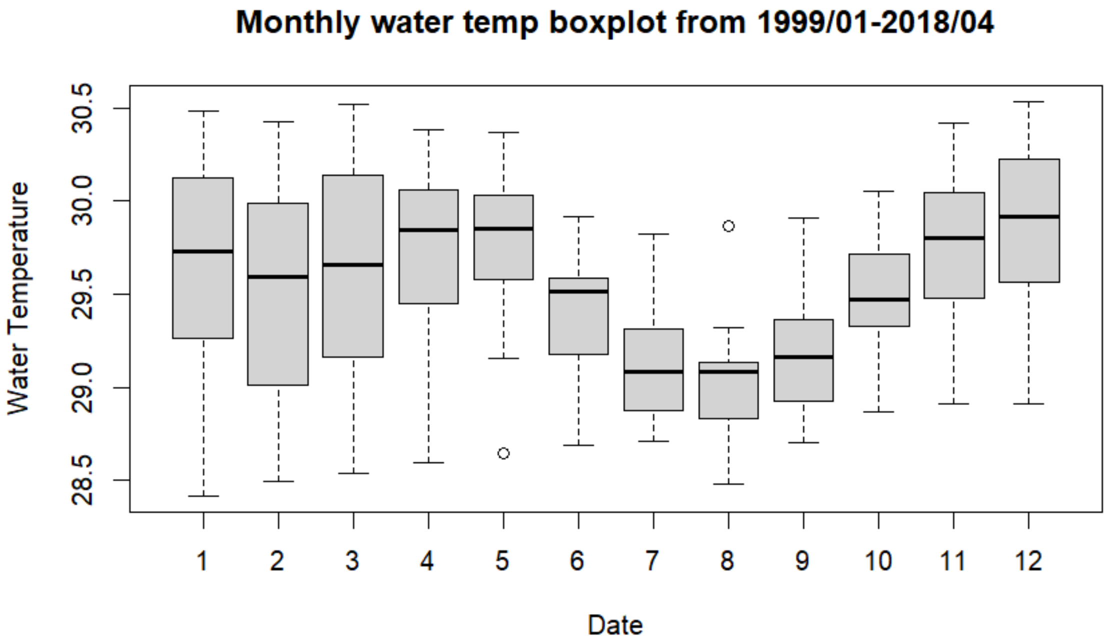


Figure 6c. The boxplot of monthly water temperature within a cycle.

It is evident that all the plots have an annual cycle. The sea level is always the highest at the beginning of the year and lowest in the middle of the year. Similarly, the air and water temperature is higher at the beginning and at the end of the year and lower in the middle of the year.

The observed phenomenon from Figure 5 and Figure 6 aligns with the seasonality of Tuvalu. There are two main seasons in Tuvalu, which are the wet season, from November to March, and the dry season, from March to November. During the wet season, Tuvalu has higher temperatures and higher precipitation, while in the dry season, Tuvalu has relatively cooler temperatures and lower precipitation. That being said, the seasonal component of the datasets can be well-explained by the wet and dry seasons of Tuvalu and therefore it doesn’t need further analysis.

STL: Trend

Since the goal of the project is to investigate the relationship between the rising temperature due to global warming and the rising sea level in Tuvalu, it is necessary to examine the relationship between the trend of temperature and the trend of sea level. The trends are then visualized and compared in Figure 7.

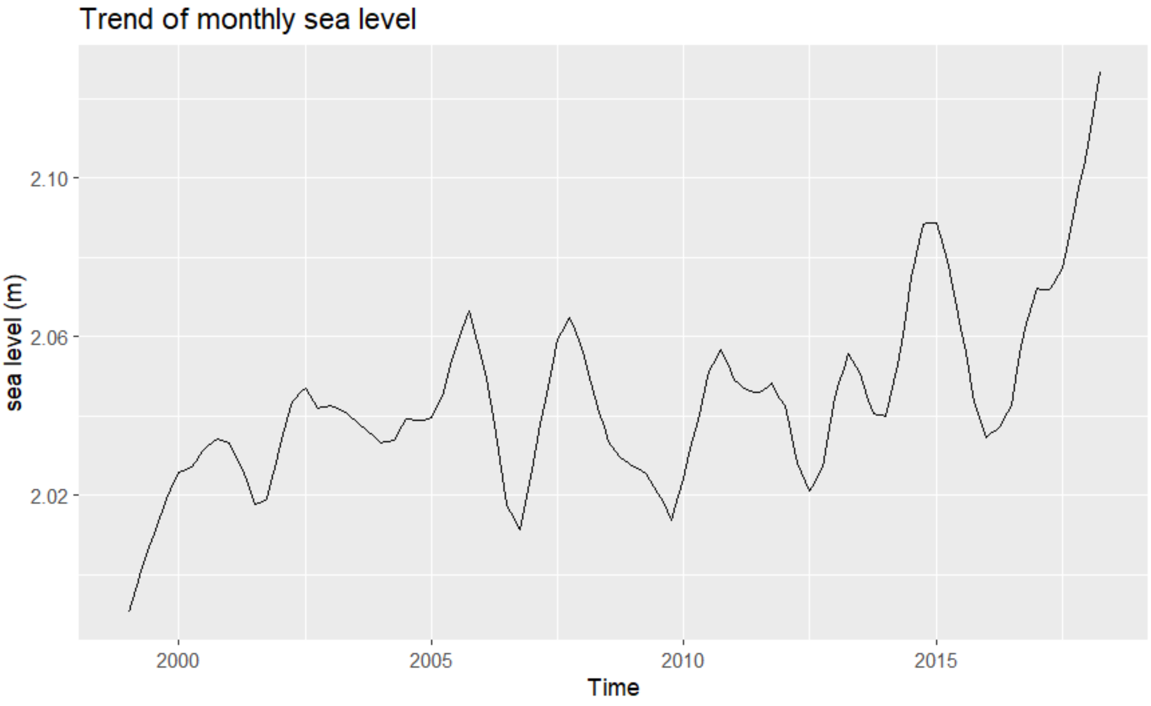


Figure 7a. The trend of monthly sea level, $T_S(t)$

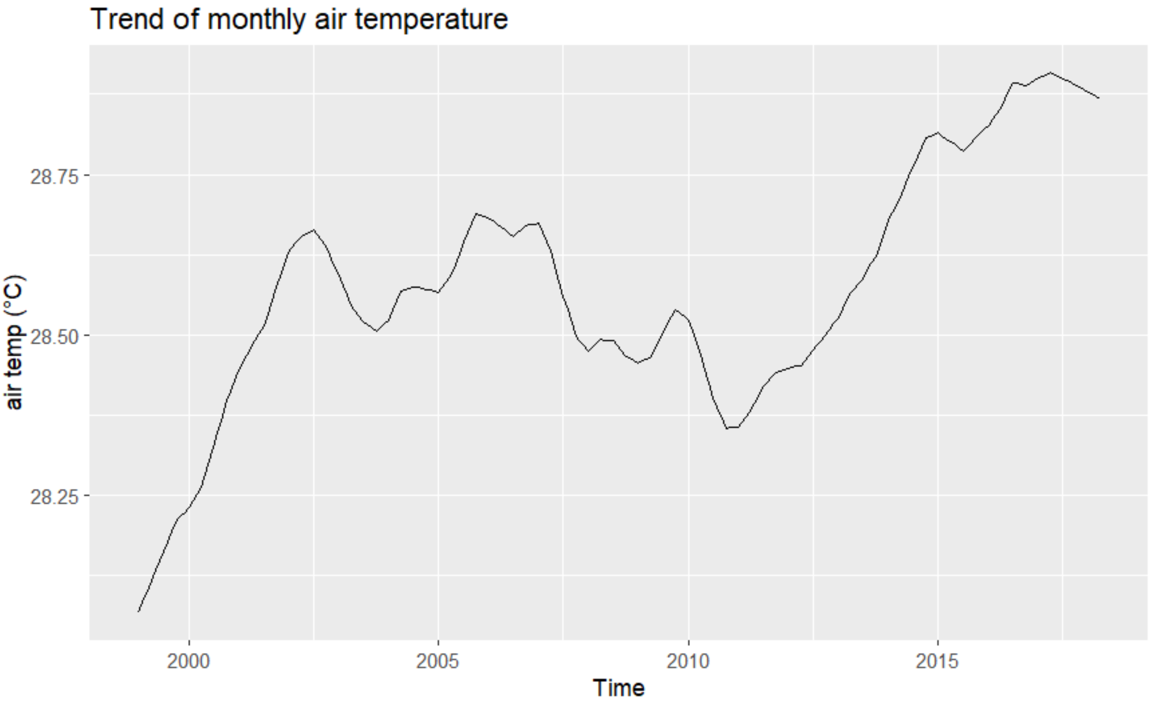


Figure 7b. The trend of monthly air temperature, $T_A(t)$

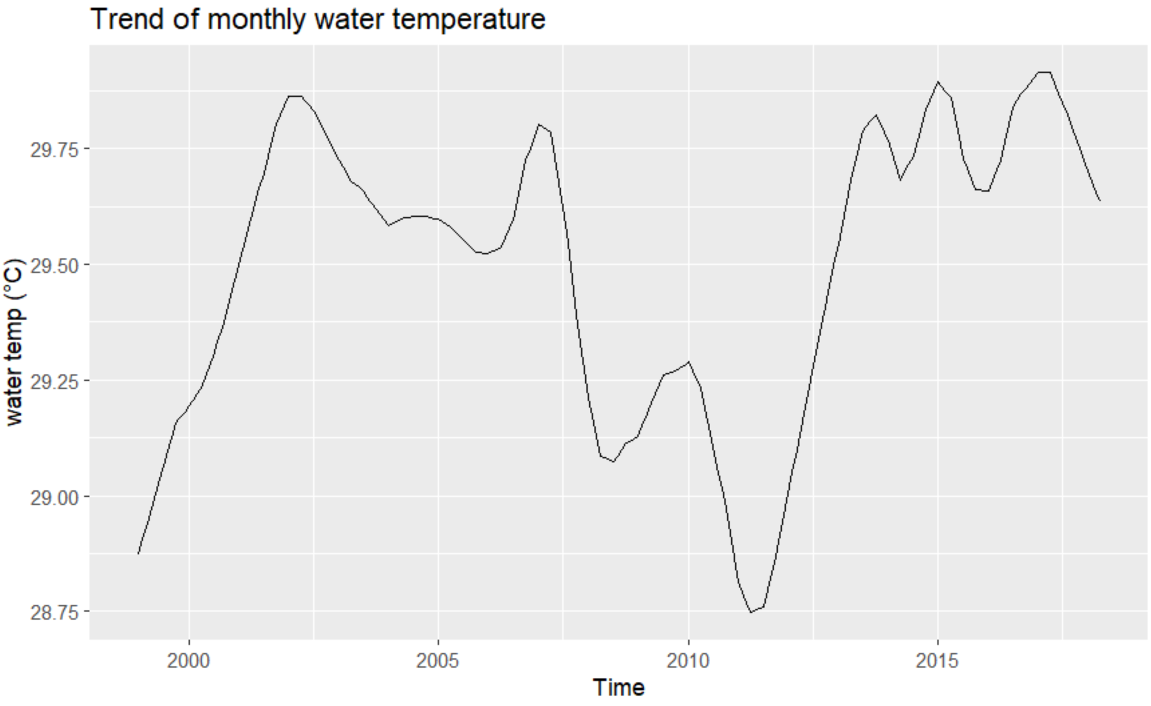


Figure 7c. The trend of monthly water temperature, $T_W(t)$

Based on the figures presented above, there is an increasing trend in the sea level, water temperature, and air temperature with the water temperature plot fluctuating significantly. Then, the next step is to develop ARIMAX models (an extension of the ARMA model with I standing for integrated and X standing for exogenous factors, here, X is the water temperature and air temperature) that computes the trend value of the sea level according to the trend value of the air temperature, or the trend of water temperature. In this way, the future sea level can be predicted as a function of the air temperature or water temperature trend which changes due to global warming.

To enhance the visibility of the trends, simple linear regressions are performed on the raw data and described as trend lines in Figure 8.

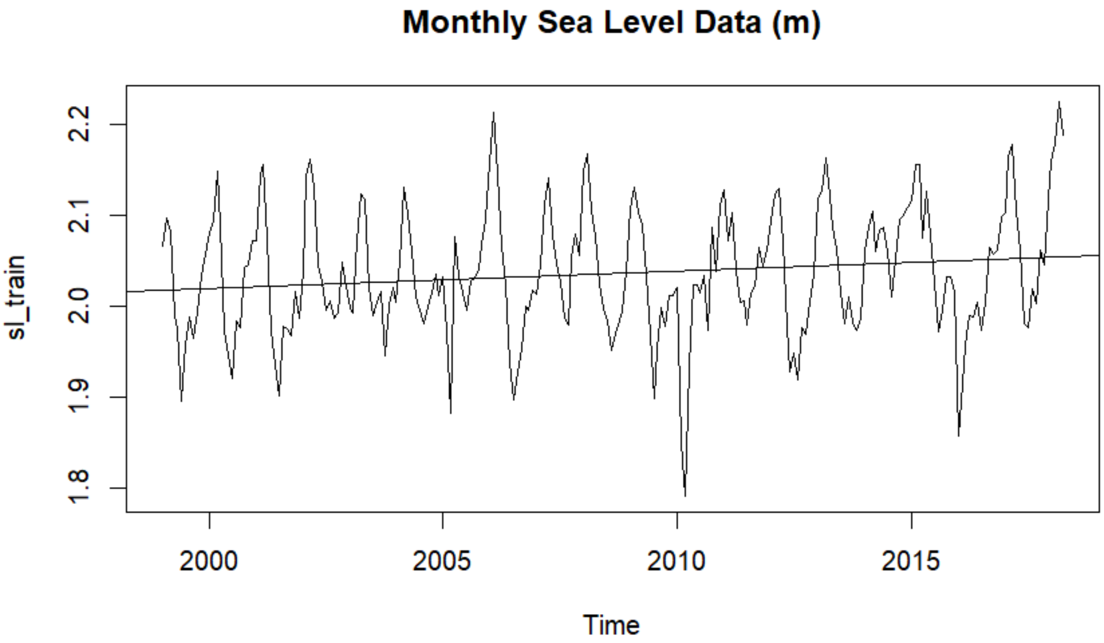


Figure 8a. The linear trend line of monthly sea level.

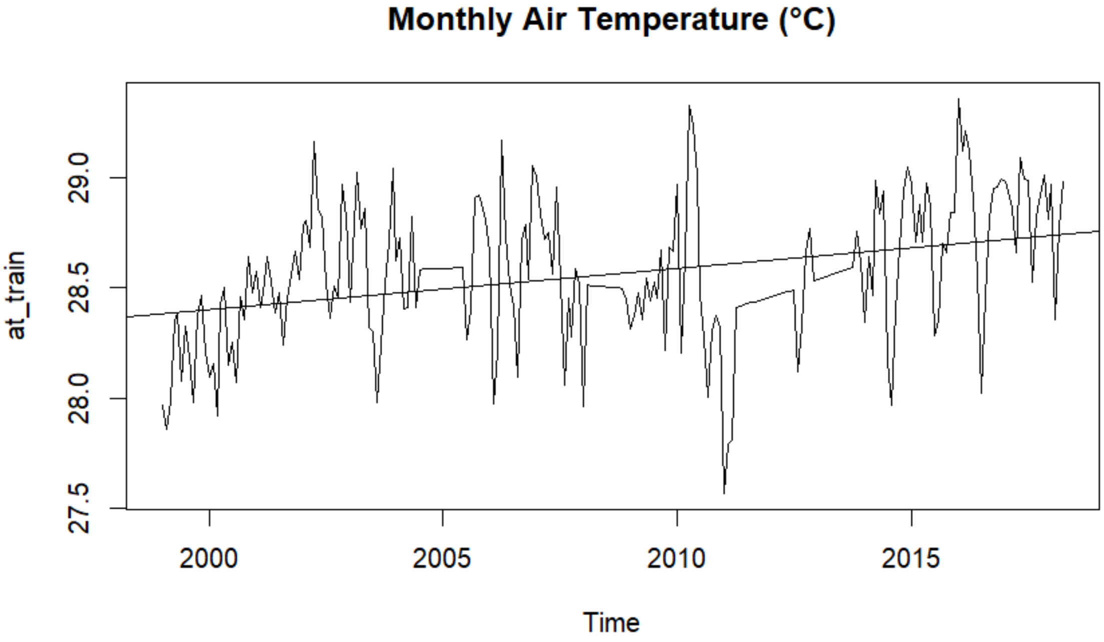


Figure 8b. The linear trend line of monthly air temperature.

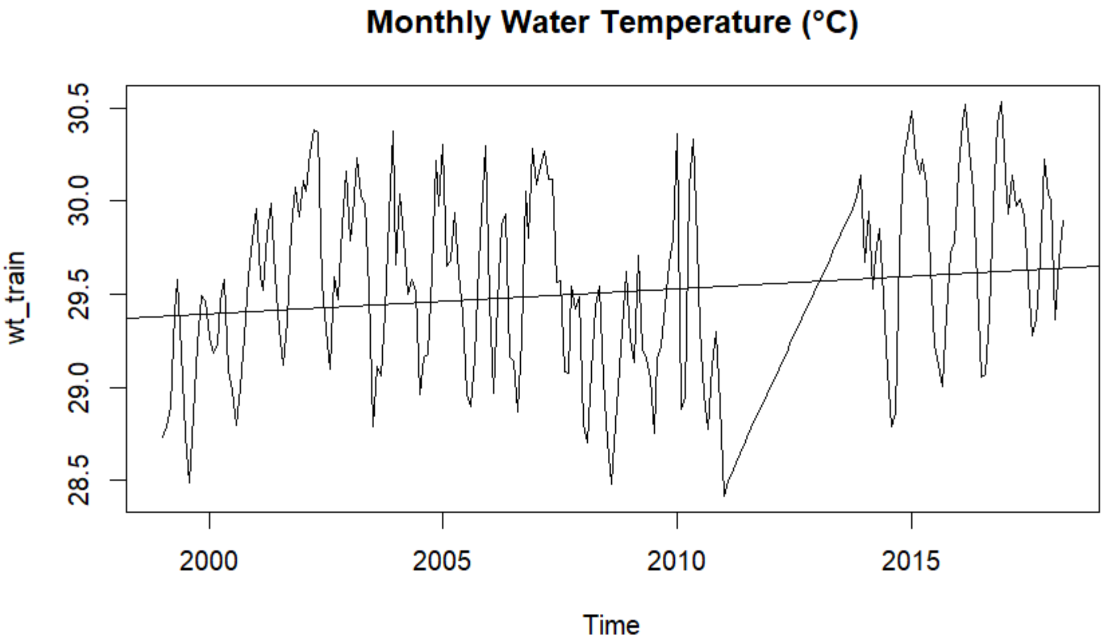


Figure 8c. The linear trend line of monthly water temperature.

It is clear that as time goes by, the sea level increases and the temperatures rise. This observation can be taken as evidence of the existence of global warming and the rise of sea levels in Tuvalu.

STL: Forecast

Hereby, the forecasts of the sea level and temperatures based on the STL model are performed to be compared with the test dataset to evaluate the performance of the model. The prediction is then made within the time period from April 2018 to February 2023 and overlapped with the test values as shown in Figure 9. The red line on the left-hand to the dashed line represents the train data. The red line on the right-hand to the dashed line represents the test values. The blue line represents the forecast values from the STL model. The gray area is the confidence interval, with the 95% confidence interval being the lighter gray and the 80% confidence interval being the darker one.

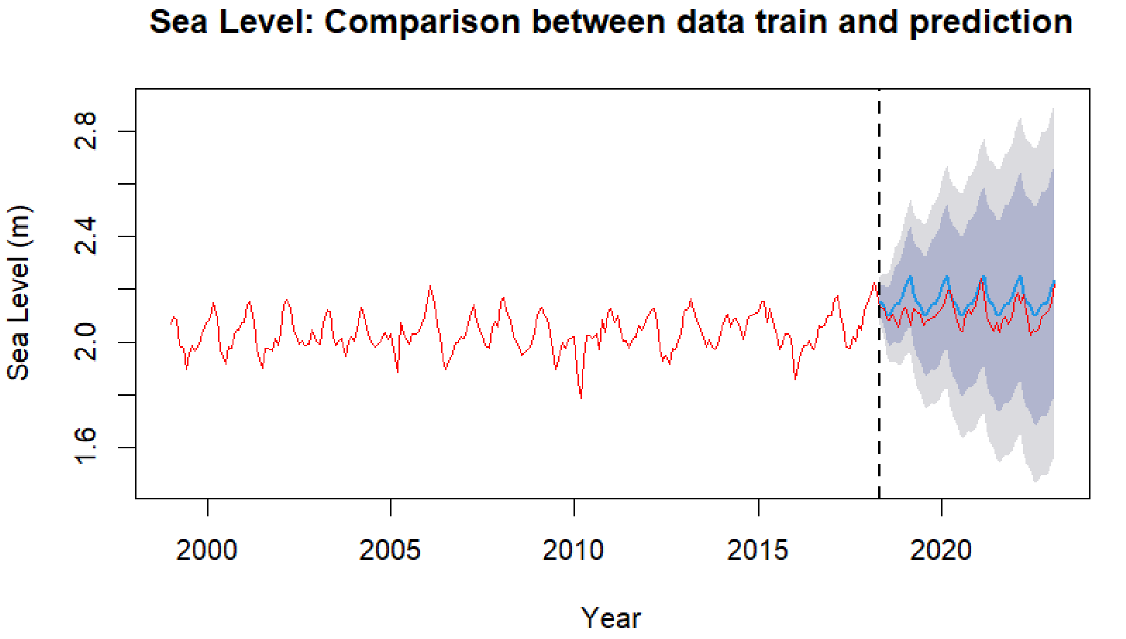


Figure 9a. The comparison of sea level between the STL model forecasting (blue) and the test data (red).

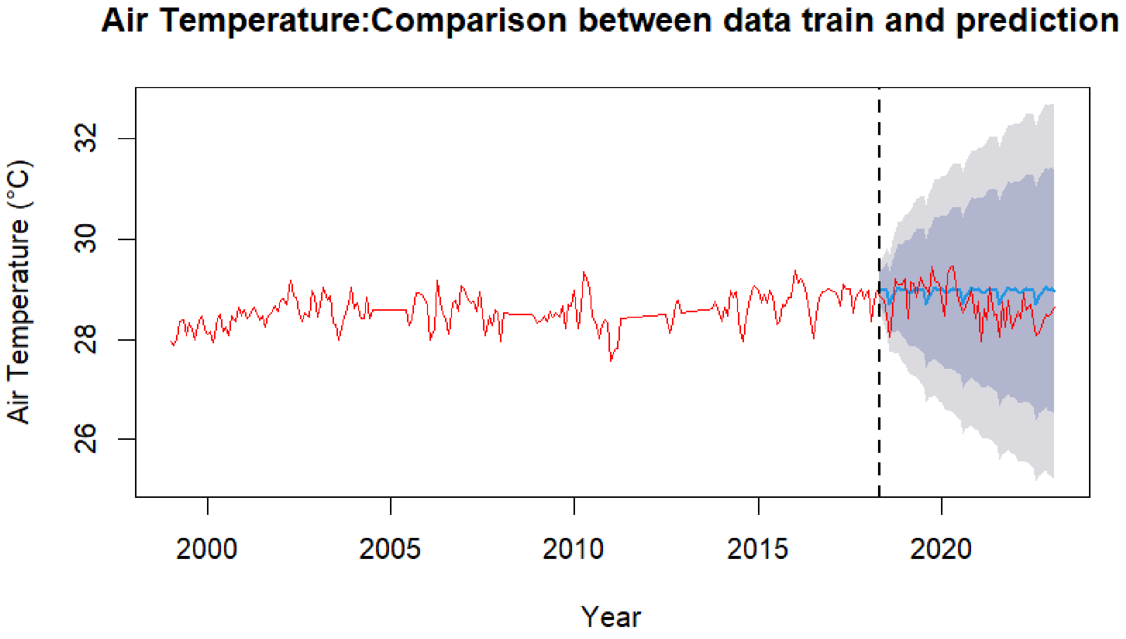


Figure 9b. The comparison of air temperature between the STL model forecasting (blue) and the test data (red).

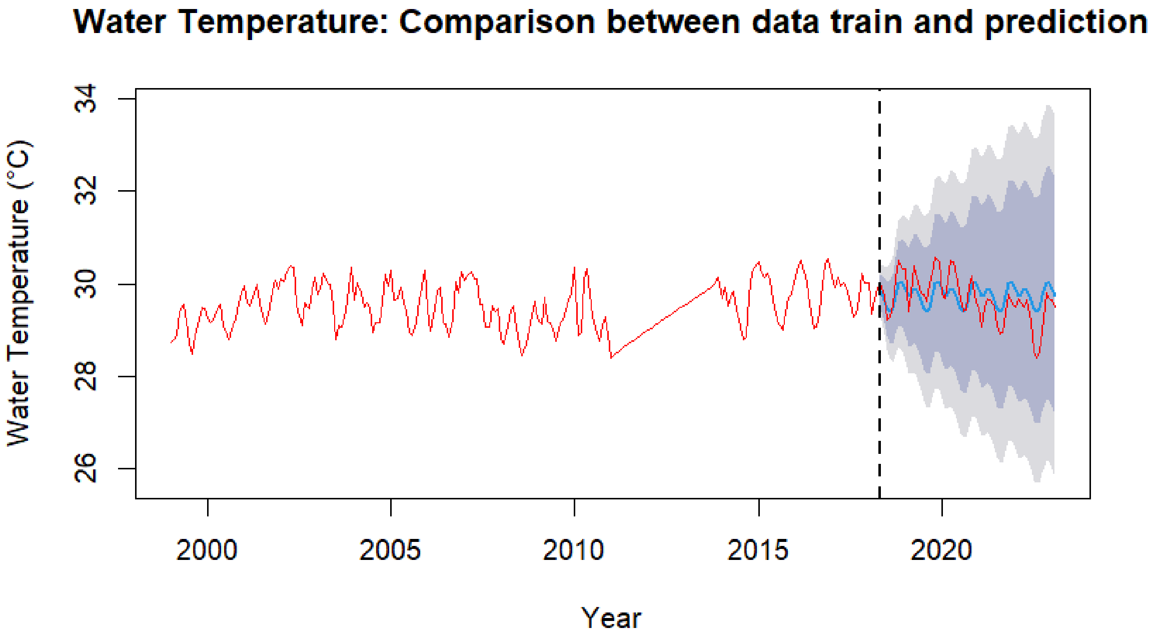


Figure 9c. The comparison of water temperature between the STL model forecasting (blue) and the test data (red).

Figure 9 shows that the predicted values closely match the test values, falling within the confidence interval. This indicates that the predictions generated using the STL model are accurate and perform well.

STL: Results

The detailed numerical results of the STL model are recorded in Table 2. The performance of the STL model is evaluated not only through Figure 9 above but also through the metrics. The first metric is the Mean Absolute Percentage Error (MAPE), measuring the percentage difference between the generated values and the actual values. The second metric is Root Mean Squared Error (RMSE), which is the square root of the average of the squared differences between the generated values and the actual values. These two metrics are commonly used to examine the accuracy of models.

Table 2. The results of the STL model

Method	Data		MAPE	RMSE
STL + Random walk	Sea level	Training	0.5200313	0.0448752
		Test	0.02456842	0.06116707
	Air Temp	Training	0.6300398	0.2509793
		Test	0.01254471	0.4198845
	Water Temp	Training	0..6614255	0.2637418
		Test	0.01161714	0.4082814

The training MAPE and RMSE are obtained through comparing the data from the fitted STL model and the actual training data. The test MAPE and RMSE are obtained through the difference between the predictions generated from the fitted STL and the test data. The results from the STL decomposition model show that all of the MAPE and RMSE values are small, which suggests a high level of accuracy in fitting the training data into the model as well as forecasting the sea level and the temperatures.

ARIMAX

In accordance with the hypothesis made in this project that the sea level is influenced by temperature change caused by global warming in the long term, the monthly sea level data is considered as a function of the temperatures. Then, as described above, the ARIMAX model is applied here while being combined with the STL decomposition model (since only trend data are needed in this model and the trends are obtained using the STL decomposition) to analyze the relationship between trends of sea level and the temperatures to improve the accuracy of the prediction of the sea level rise in Tuvalu. The ARIMAX model is thus a predictive model that analyzes past performances to make future predictions. With the model, the project can utilize the effects of global warming to make a precise prediction of the future sea level in Tuvalu than making predictions only based on the STL model.

Three types of models are established to analyze the relationships. The first model is the trend of sea level as a function of the trend of water temperature, the second model is the trend of sea level as a function of the trend of air temperature, and the third model is the trend of sea level as a function of the combination of the trends of air and water temperatures as a matrix.

Worth mentioning, there are three kinds of orders in the ARIMAX model, which are represented as p, d, and q. The first notation p stands for the order of the AR component, that is the coefficient of the model that describes the relationship between an observation and a linear combination of past observations. The second notation d stands for the order of differencing, which is the number of times the time series data are differenced to make the data stationary. The third notation q stands for the order of the MA component, which is the coefficient of the model that describes the relationship between an observation and the error terms associated with the past observations. Thereby, cross-validation is used to optimize the three ARIMAX relationship models by finding the best $p,\ d,$ and q orders for each model.

ARIMAX: Cross-validation

To determine the appropriate ARIMAX orders for the sea level trend data from STL decomposition, different models with different orders are built and compared for the best performance. A loop is built to build all ARIMAX models with p ranging from 0 to 5, d ranging from 0 to 1, and q ranging from 0 to 5 (Zvornicanin 2022). As a result, 72 candidate ARIMAX models are built for each relationship.

Then, for each of the 72 candidate ARIMAX orders, a Rolling-window Time Series Cross Validation method is employed for the model validation. This cross-validation method performs a walk-forward validation on the sea level trend data, where each validation set consists of a single observation, and the corresponding training set includes only observations that occurred prior to the validation set. The forecast accuracy is computed by averaging the validation sets (Keith 2022). The mechanism of the cross-validation is visualized in Figure 10, where blue observations represent the training sets, and orange observations represent the validation sets.

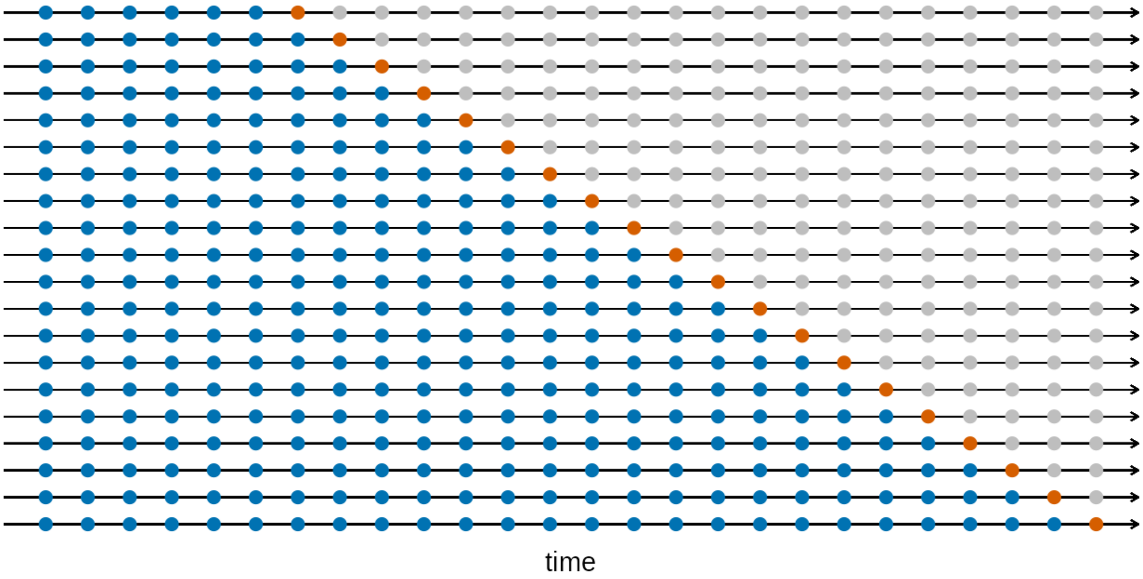


Figure 10. The Rolling-window Time Series Cross-validation (Hyndman 2021)

The performance of the candidate models is evaluated with metrics MAPE and RMSE. The best models are chosen based on these metrics. Due to the limited length of the paper, the detailed performance of the 72×3 models is described in Appendix A.

In the following paper, trend of sea level is expressed as T_s, while trend of water temperature is T_w, and trend of air temperature is T_a.

Figure 11 visualizes the rolling-window cross-validation results.

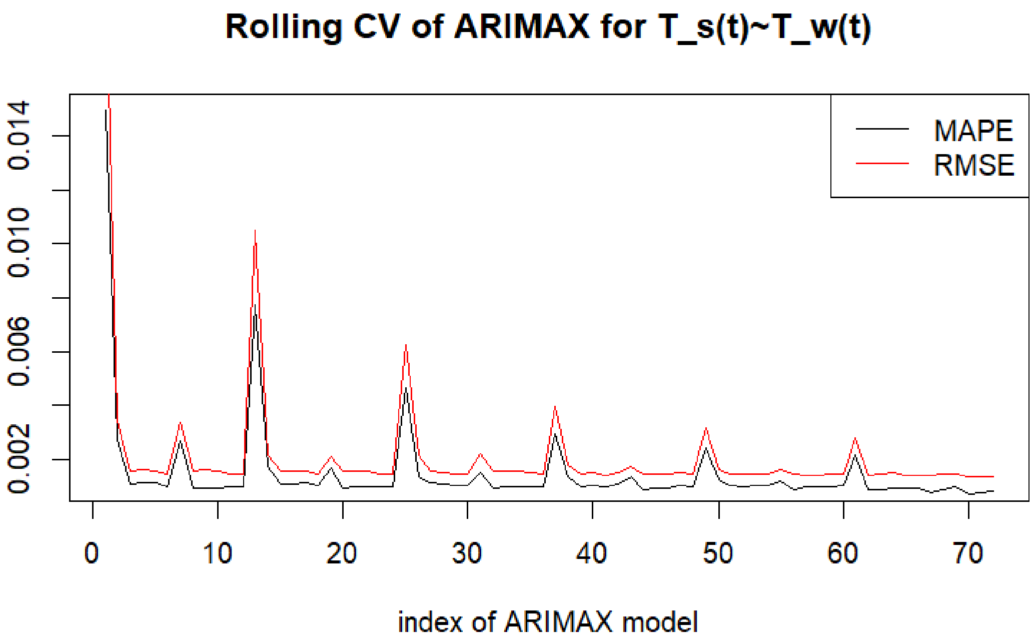


Figure 11a. $T_s(t) \sim T_w(t)$ ARIMAX rolling-window cross-validation results.

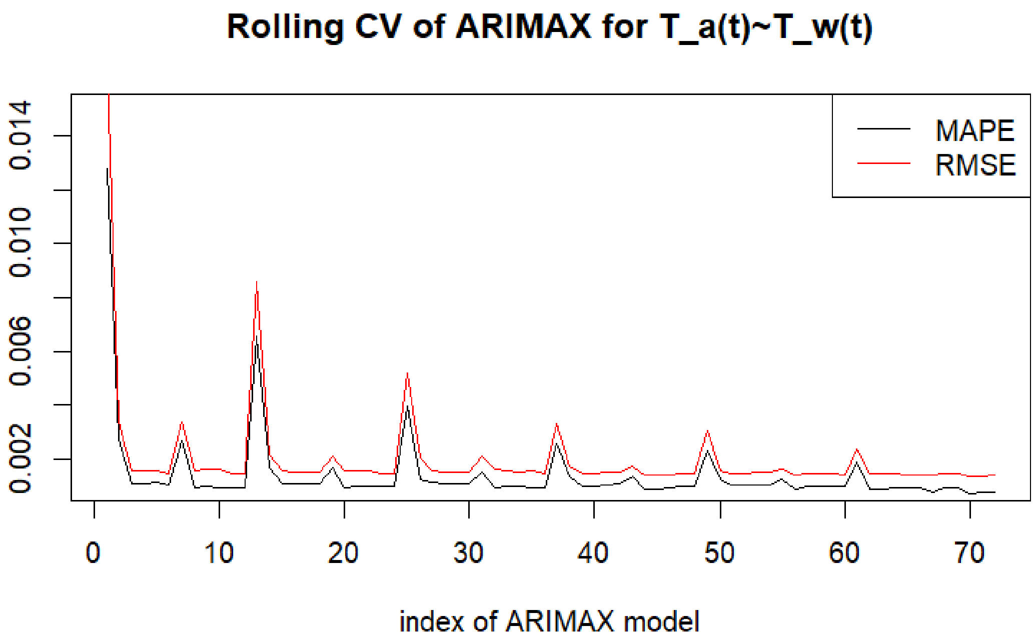


Figure 11b. $T_s(t) \sim T_a(t)$ ARIMAX rolling-window cross-validation results.

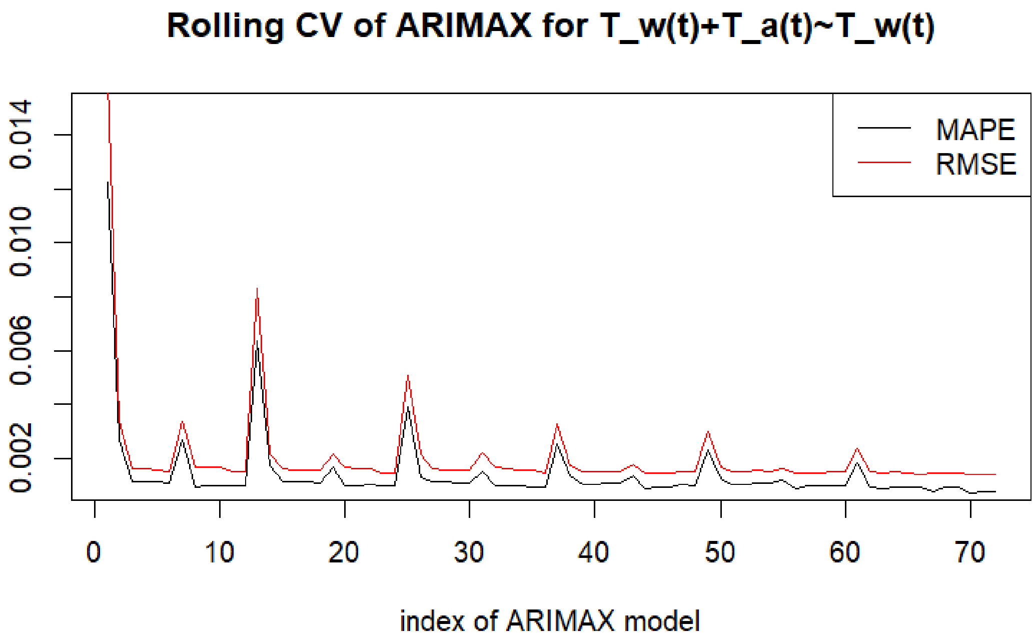


Figure 11c. $T_s(t) \sim T_w(t)\ +T_a(t)$ ARIMAX rolling-window cross validation results.

From Figure 11, it can be seen that both MAPE and RMSE are the lowest at the 70th candidate model for all three ARIMAX relationship models, which is the ARIMAX(3,1,5) model, while the numerical results in Appendix A support the conclusion. The best orders for all three relationship models are therefore ARIMAX(3,1,5) as a result.

ARIMAX: Results

After combining the ARIMAX model with the trend components of the STL model to predict the future sea level based on the air temperature, Figure 12 is generated.

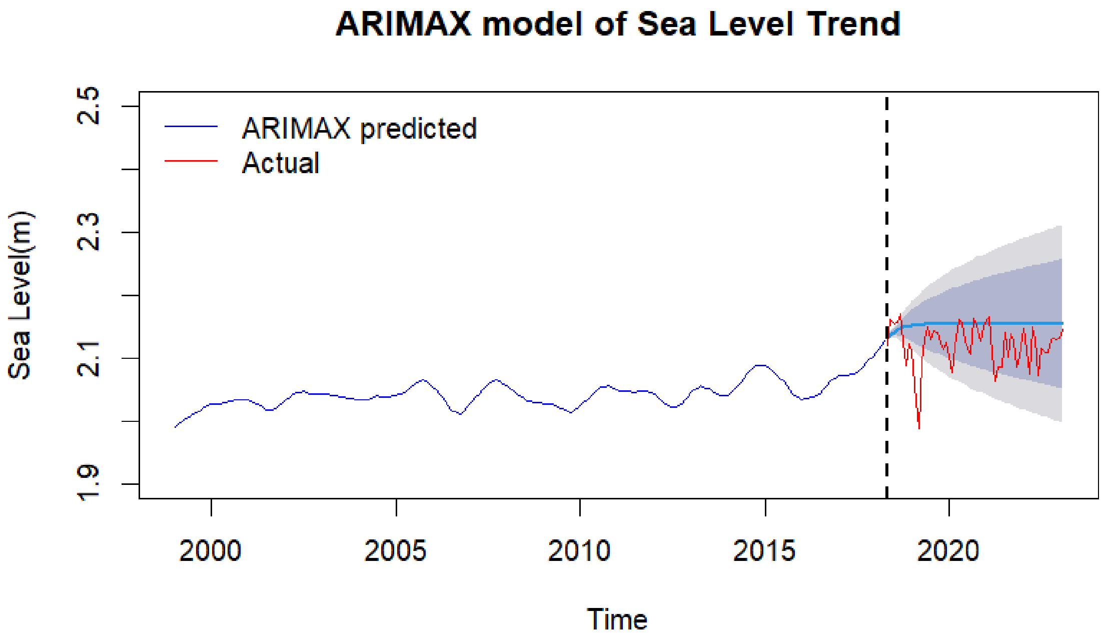


Figure 12a. $T_s(t) \sim T_w(t)$ ARIMAX(3,1,5) model without seasonal and residual components.

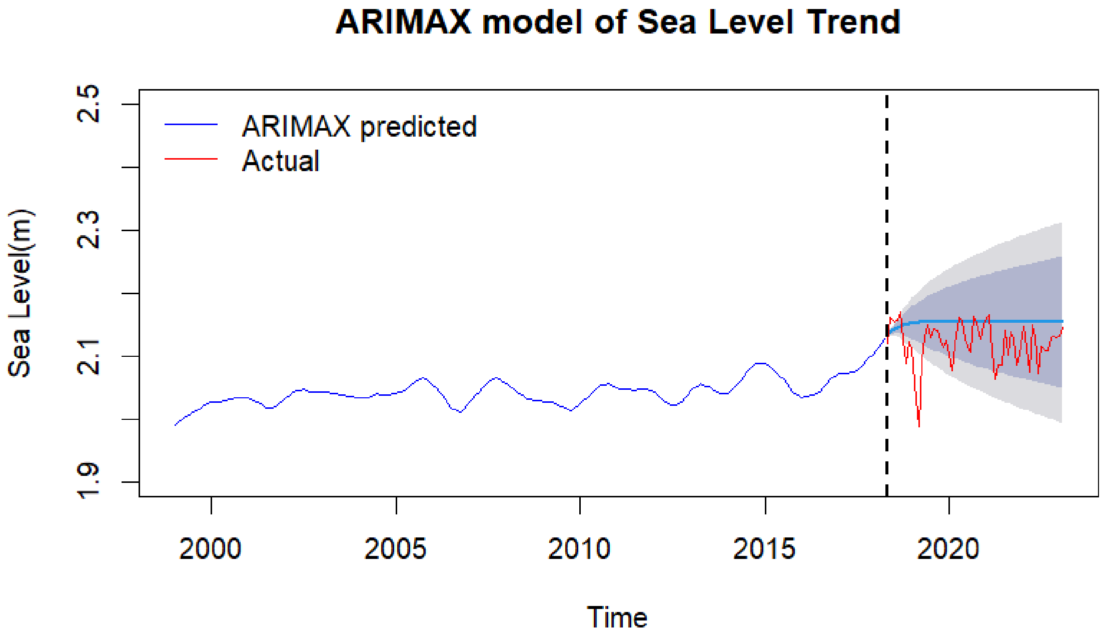


Figure 12b. $T_s(t) \sim T_a(t)$ ARIMAX(3,1,5) model without seasonal and residual components.

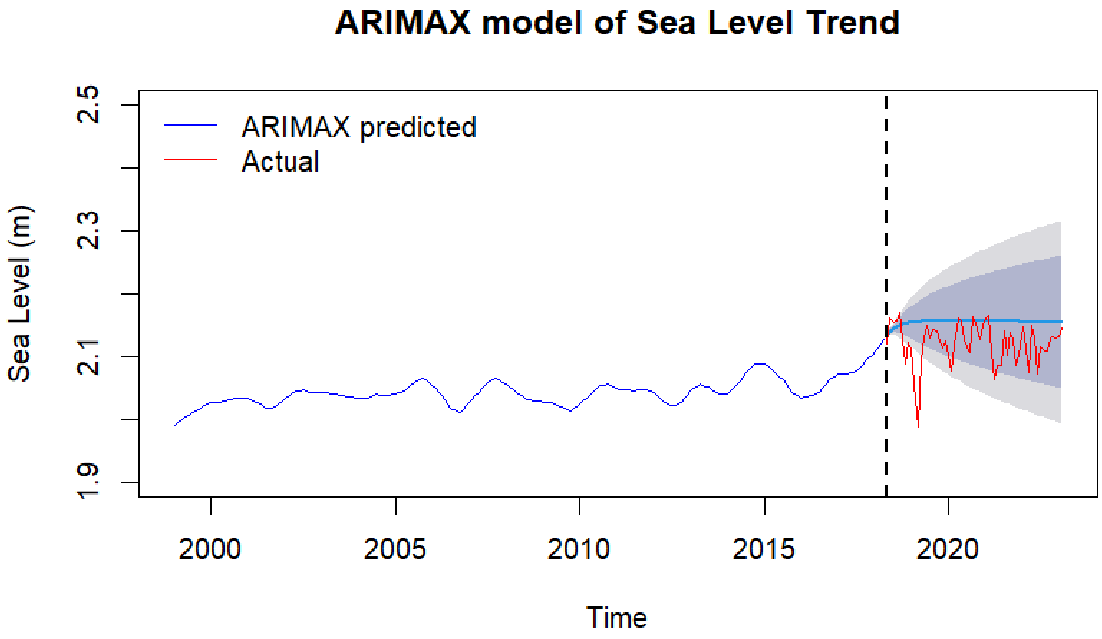


Figure 12c. $T_s(t) \sim T_a(t) + T_w(t)$ ARIMAX(3,1,5) model without seasonal and residual components.

As mentioned in the previous section, the blue line represents the forecasting of the sea level while the red line stands for the actual sea level. The gray area represents the confidence interval with the 80% confidence interval being the darker one and the 95% confidence interval being the lighter area. The actual sea level data exceeds the 95% confidence interval while the forecasting sea level matches the general trend of the actual sea level data. The model performance is hence not well and needs to be further improved.

Then, seasonality and residual part are added to the trend generated by the ARIMAX model as shown in Figure 13.

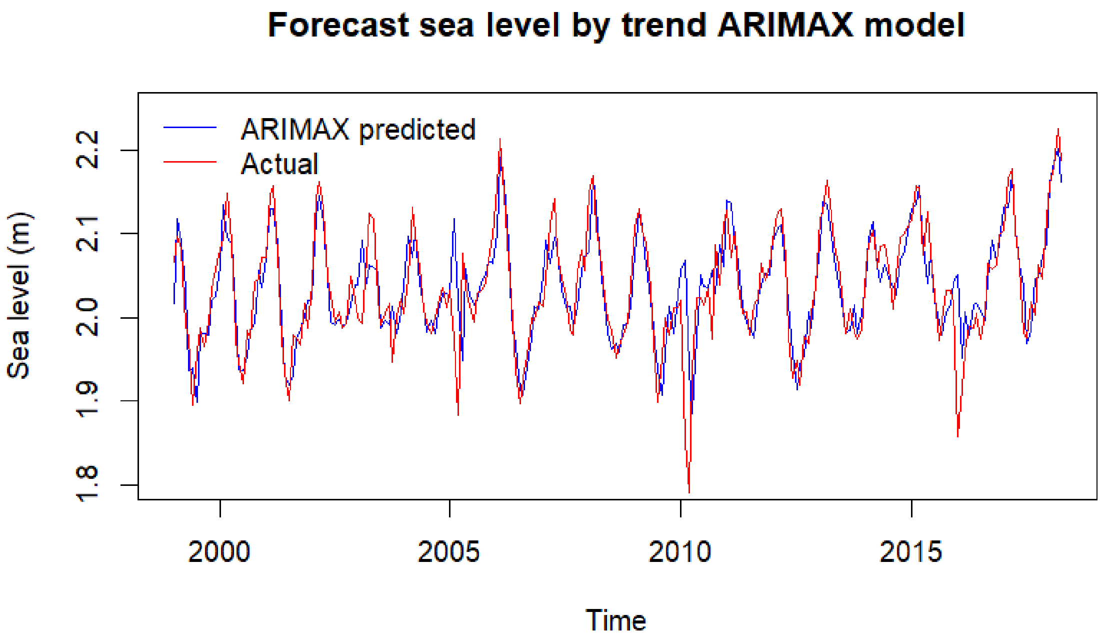


Figure 13a. $T_s(t) \sim T_w(t)$ ARIMAX(3,1,5) model with seasonal and residual components

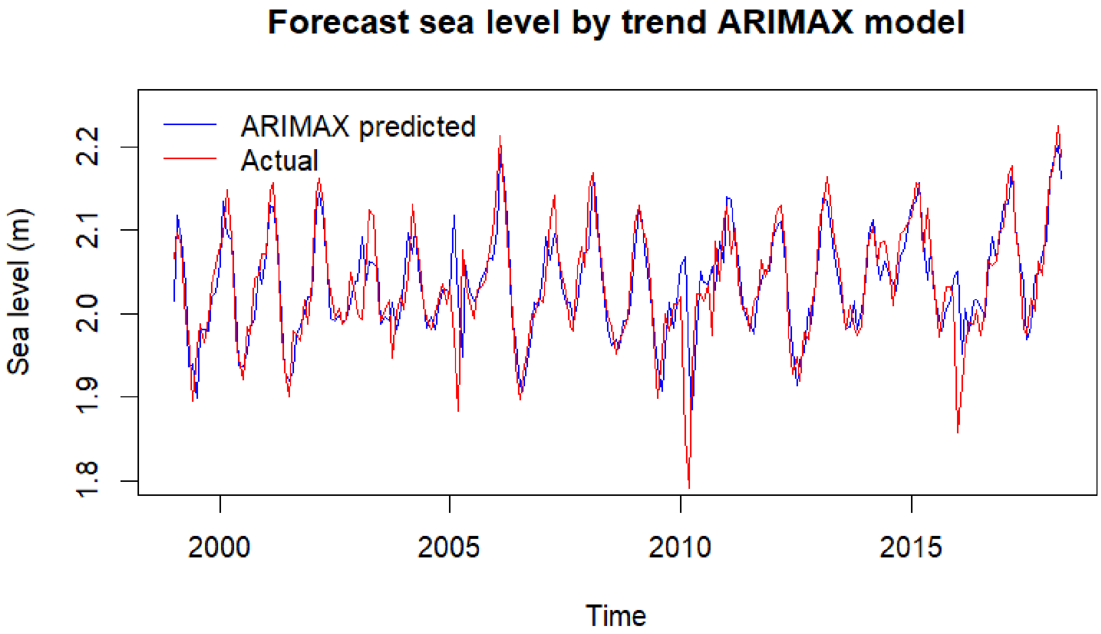


Figure 13b. $T_s(t) \sim T_a(t)$ ARIMAX(3,1,5) model with seasonal and residuals components

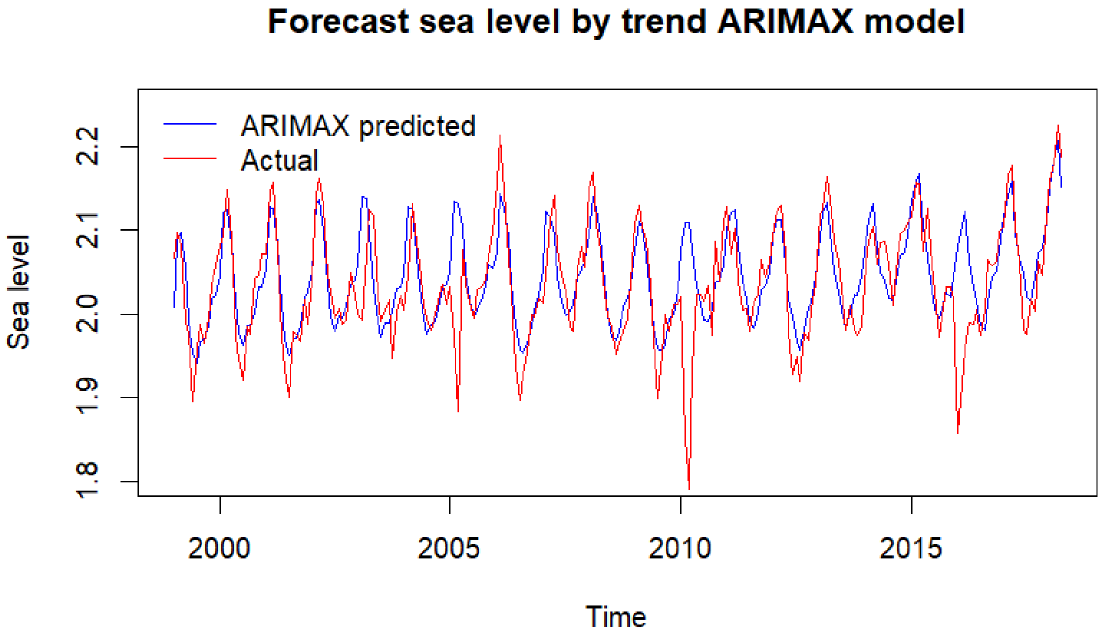


Figure 13c. $T_s(t) \sim T_w(t) \ + T_a(t)$ ARIMAX(3,1,5) model with seasonal and residual components

In Figure 13, the prediction is drawn in blue and the actual data is drawn in red. In all three models, the blue lines are close to the red lines, indicating the models fit well with the addition of seasonal components. The three models are used for the final predictions.

MLM

MLM refers to Machine Learning Morphism, which is a building block for designing machine learning workflows from the very beginning of data cleaning to the end of the project completion of the task. Specifically, a MLM can be described as 5 tuples as listed in the 9 MLM models below, which are the input space, output space, learning morphism, prior parameters, and risk function, while the models are data cleaning, Kalman filter, ADF test, LOESS model, STL decomposition model, ARMA model, ARIMAX model, rolling-window validation, model performance, and linear regression which is used for the final prediction (Cawi et al. 2019). In MLM, the input and output space are vector spaces that describe the input and output, the learning morphism is the function that converts the input to output, the prior parameters is the prior knowledge of the parameters in the morphism, and the risk function is the empirical risk function that is used to minimize the parameters. Therefore, MLM is a framework that organizes different models and optimizes the parameters in the process.

MLM: Models

ML_0: Data Cleaning

- Input probability space: $\mathbb{X}_0 \subseteq \mathbb{R}^m$, where $\mathbb{X}_0$ is the time series dataset includes the raw training dataset and the raw testing dataset. Here, it is assumed $x = (a, b),\ x \in \mathbb{X}_0$, a is the time part in months format, while b is the value part and is a real value.
- Output probability space: $\mathbb{Y}_0 = \mathbb{R}^m$, where $\mathbb{Y}_0$ is the time series dataset includes the training dataset and the testing dataset without null data and El Niño effects.
- Learning morphism:
  - Cleaning null data: when the value part of a data point in $\mathbb{X}_0$ doesn't exist, omit the data point; in other words, when $F_{ML_{0-1}}: \mathbb{Y}_0 = \{x(a, b) | x \in \mathbb{X}_0,\ a \in date,\ b \in \mathbb{R},\ b \neq "NA"\}$
  - Removing outliers (strong El Niño effects): when the time part of a data point in $\mathbb{X}_0$ is included in the strong El Niño effect period, omit the data point; in other words, $F_{ML_{0-2}}: \mathbb{Y}_0 = \{x(a, b) | x \in \mathbb{X}_0,\ a \in date,\ b \in \mathbb{R},\ a \notin ["1997-03", "1998-12"]\}$
- Prior distribution on parameters:
  - Cleaning null data: $b = "NA"$
  - Removing outliers: $a \in ["1997-03", "1998-12"]$

- Empirical risk function: $E(X - Y)$, $X \in \mathbb{X}_0 \cap \mathbb{Y}_0$, $Y \in \mathbb{Y}_0$, X and Y are in the same month but different years, representing the risk of eliminating normal data as outliers.

ML_1: Kalman Filter

- Input probability space: $\mathbb{X}_1 \subseteq \mathbb{R}^m$, where $\mathbb{X}_0 \cup \mathbb{Y}_0 = \mathbb{X}_1$, $\mathbb{X}_0$ and $\mathbb{Y}_0$ are the time series input and output from ML_0.
- Output probability space: $\mathbb{Y}_1 = \mathbb{R}^m$, a time series dataset, with $\mathbb{Y}_1$ including $\mathbb{X}_1$ and the filled data points generated by F_{ML_1}
- Learning morphism:
  - Detecting the missing data: $\{X | X \in \mathbb{X}_0, X \notin \mathbb{Y}_0\}$;
    Finding the observed data: $\{Y | Y \in \mathbb{Y}_0\}$;
    Making prediction of the missing data: $\widehat{X}$ based on Y
  - Finding the value for the missing data by prediction:
    $F_{ML_1}: \widehat{x_{t+1}} = F\widehat{x_t} + \Theta_t \Delta_t^{-1}(y_t - G\widehat{x_t})$; $\Omega_{t+1} = F\Omega_t F^T + \Sigma_V - \Theta_t \Delta_t^{-1} \Theta_t^T$,
- Prior distribution on parameters:
  - $x_t \in X, y_t \in Y, \widehat{x_t}$ is based on y_n, $1 \leq n \leq t - 1$
  - Information for the latest observation y_t: $I_t = y_t - \hat{E}_{t-1}(y_t)$;
    $E\{I_t\} = 0$, I_t is independent of x_t, orthogonal to all linear combinations of $y_{1, \dots, t-1}$ and orthogonal to $I_{t-1, t-2, \dots}$
  - Prediction operator: $\hat{E}_{t-1}(y_t) = G\widehat{x_t}$, $I_t = G(x_t - \widehat{x_t}) + w_t$, $w_t \in W$ is the white noise series, independent of y_t, x_t *and* $\widehat{x_t}$
  - $\hat{E}_{t-1}(x_{t+1}) = \hat{E}_{t-1}(Fx_t + V_{t+1}) = F\widehat{x_t}$, V_{t+1} is independent of x_t, $\widehat{x_t}$, and w_t;
  - $\Delta_t = G\Omega_t G^T + \Sigma_w$, Δ_t is symmetric
  - $\Theta_t = F\Omega_t G^T = E\{(Fx_t + V_{t+1})(x_t - \widehat{x_t})^T G^T + W_t^T\} = E\{x_{t+1} I_t^T\}$
- Empirical risk function: $\Omega_t = E\{(x_t - \widehat{x_t})(x_t - \widehat{x_t})^T\}$, representing the prediction quadratic error.

ML_2: ADF Test

- Input probability space: $\mathbb{X}_2 = \mathbb{Y}_1 \subseteq \mathbb{R}^m$. It is a time series dataset and the output from ML_1.

- Output probability space: $\mathbb{Y}_2 \subseteq (0, 1]$.
- Learning morphism: F_{ML_2}: Finding y: $x_t = \mu + yx_{t-1} + w_t$, $x \in \mathbb{X}_2$, $y \in \mathbb{Y}_2$; x_{t-1} means lag 1 of the series
  - When $y = 1$, there is a unit root, $x \in \mathbb{X}_2$ is non-stationary, the null hypothesis can not be rejected.
  - When $y < 1$, there is no unit roots, $x \in \mathbb{X}_2$ is stationary, the null hypothesis can be rejected.
- Prior distribution on parameters: μ is a constant that can be 0, when it is 0; $w_t \in W$ is the white noise so it is a sequence of independent identically distributed random variables that can be assumed to have a Gaussian distribution.
- Empirical risk function: the test statistic $DF = \frac{\hat{y}}{\widehat{\sigma^2}}$, where $\hat{y} = \frac{\Sigma x_i x_{i-1}}{\Sigma x_{i-1}^2}$, is the least squares estimate of y, and $\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{y}x_{i-1})^2$, is the variance of w_t. DF is compared to the relevant critical value of the ADF test.
  - When test statistic is equal to or larger than the critical value, there is a unit roots, $x \in \mathbb{X}_2$ is non-stationary, the null hypothesis can not be rejected.
  - When test statistic is smaller than the critical value, there is no unit roots, $x \in \mathbb{X}_2$ is stationary, the null hypothesis can be rejected.
  - There can be type I or type II error, when $y = 1$ while the test statistic is smaller than the critical value for the ADF test, or, when $y < 1$, but the test statistic is larger than or equal to the critical value.

ML_3: LOESS Model

- Input probability space: $\mathbb{X}_3 \subseteq \mathbb{R}^m$. It is a time series dataset.
- Output probability space: $\mathbb{Y}_3 \subseteq \mathbb{R}^m$ and is a time series dataset.
- Learning morphism: F_{ML_3}: the weighted regression $Y = W^T X$, $X \in \mathbb{X}_3$, $Y \in \mathbb{Y}_3$, W is the weight.
- Prior distribution on parameters: W is a $m \times m$ matrix with entries being $W_t = (1 - d_t)^3$ and d_t is the distance between $\hat{x}_t$ and x_t
- Empirical risk function: the least squares error $(Y - W^T Y) \cdot W \cdot (Y - W^T X)$

ML_4: STL Decomposition Model

- Input probability space: $\mathbb{X}_4 \subseteq \mathbb{Y}_1 \subseteq \mathbb{R}^m$, and is the training dataset part of the output from ML_1.
- Output probability space: $\mathbb{Y}_4 \subseteq \mathbb{R}^m$ and has 3 time series datasets, which are $\{T,\ S,\ R\}$, and each of them $\subseteq \mathbb{R}^m$.
- Learning morphism: $F_{ML_4}:\ T_t + S_t + R_t = X_t$, T is the trend component, S is the seasonal component, R is the residual component, $X \in \mathbb{X}_4$
    - ○ Step 1: Using LOESS Model as described in ML_3 to fit the detrended component to obtain the cycle-subseries smoothing component: $C_t^{k+1} = F_{ML_3}(x_t - T_t^{k})$
    - ○ Step 2: Using LOESS Model and two MA filters to obtain the low-pass filtering component of smoothed cycle-cubseries L_t^{k+1}
    - ○ Step 3: Finding the seasonal component: $S_t^{k+1} = C_t^{k+1} - L_t^{k+1}$
    - ○ Step 4: Using LOESS Model to update the trend component: $T_t^{k+1} = F_{ML_3}(x_t - S_t^{k+1})$
    - ○ Step 5: Obtaining the residual component: $R_t^{k+1} = x_t - S_t^{k+1} - T_t^{k+1}$
- Prior distribution on parameters:
    - ○ In the first iteration, the trend component $T_t^{0} = 0$
    - ○ The Random Walk model is used to fit the residual component R, where $R_{t+1} = R_t + W_{t+1} + \mu$, W is white noise with Gaussian distribution, and μ is a constant representing a drift.
- Empirical risk function: the auto-correlation function (ACF) of the residual component R_t: $\rho(h) = corr(R_{t+h}, R_t)$, where $h \in [1, n)$; if the results of ACF are not close to 0 at all lags, then auto-correlation exists in the residual component of the STL decomposition is not a white noise (i.i.d), and ARMA in ML_5 should be applied to the residual component to separate the auto-correlation part from the residual component.

ML_5: ARMA Model

- Input probability space: $\mathbb{X}_5 \subseteq \mathbb{R}^m$, it is a time series and is equal to the residual part R of the output from ML_4 (STL decomposition).
- Output probability space: $\mathbb{Y}_5 \subseteq \mathbb{R}^m$, it is a time series, and is the predicted value of the input $x_t \in \mathbb{X}_5$ at each t using F_{ML_5}.

- Learning morphism: F_{ML_5}:

  $x_t = \mu + \sum_{i=1}^{p} \phi_i x_{t-i} + w_t + \sum_{i=1}^{q} \theta_i w_{t-i} \rightarrow y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + w_t + \sum_{i=1}^{q} \theta_i w_{t-i}$, $x_t \in \mathbb{X}_5$, $y_t \in \mathbb{Y}_5$, where it is a combination of the $AR(p)$ and $MA(q)$ models with ϕ being the AR coefficient and θ being the MA coefficient
- Prior distribution on parameters:
  - μ is a constant, $w_t \in W$ is a white noise series with Gaussian distribution, ϕ, $\theta \in \mathbb{R}$ are real numbers, p and q are positive integers.
  - x_t should be stationary and passed ML_2 (the ADF test).
- Empirical risk function: Ljung-Box Test $Q = n(n + 2) \sum_{k=1}^{h} \frac{p_k^2}{n-k}$, where n is the sample size, p_k is the sample correlation at lag k, and h is the number of lags that have been tested. Q is the test statistic of the Ljung-Box Test, when $Q > \mathcal{X}^2_{1-\alpha, h}$, α being the significant level and is usually $\alpha = 0.05$, the null hypothesis is rejected, the residual of the ARMA model is not independently distributed, indicating the presence of autocorrelation in the residual, and the residual of the ARMA model should be further fed into another ARMA model.

ML_6: ARIMAX Model

- Input probability space: $\mathbb{X}_6 \subseteq \mathbb{R}^m$; it includes two time series X_1, X_2, specifically, X_1 is the trend of the monthly sea level data from ML_4, and X_2 is the trend of the monthly water temperature data or air temperature data or the combination of these temperatures.
- Output probability space: $\mathbb{Y}_6 \subseteq \mathbb{R}^m$ which includes a time series Y, being the predicted trend of the monthly sea level data as a function of X_2
- Learning morphism: F_{ML_6}: $x_{1;t} = \mu + \sum_{i=1}^{p} \phi_i x_{1;t-i} + w_t + \sum_{i=1}^{q} \theta_i w_{t-i} + \sum_{i=1}^{r} \beta_i x_{2;t}$

  $\rightarrow y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + w_t + \sum_{i=1}^{q} \theta_i w_{t-i} + \sum_{i=1}^{r} \beta_i x_{2;t}$
- Prior distribution on parameters: similarly as in ARMA, μ is a constant, $w_t \in W$ is a white noise series with Gaussian distribution, ϕ, θ, $\beta \in \mathbb{R}$ are real numbers, p and q are positive integers.
- Empirical risk function: MAPE $\frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - x_{1;i}|}{|x_{1;i}|} \times 100\%$, when MAPE is smaller, the performance of the model is better, and the model is chosen.

ML_7: Forward Rolling-window

- Input probability space: $\mathbb{X}_7 \subseteq \mathbb{Y}_6 \subseteq \mathbb{R}^m$; $\mathbb{X}_7$ is the trend part of the monthly sea level data from ML_4
- Output probability space: $\mathbb{Y}_7 \subseteq \mathbb{R}^m$, which is the predicted time series.
- Learning morphism: F_{ML_7}: $y_t = F_{ML_6}(x_{t-k+1},..., x_t)$, where $x \in \mathbb{X}_8$, y is the output from ML_6 when the input is a time series from x_{t-k+1} to x_t.
- Prior distribution on parameters: $k = 58$ in the first model, and it increases by 1 in every model; $t = 58$ in the first model, and also increases by 1 in every model.
- Empirical risk function: MAPE $\frac{1}{n}\sum_{i=58}^{n}\frac{|x_i - y_i|}{|x_i|} \times 100\%$, when MAPE is smaller, the performance of the model is better, and the model will be chosen.

ML_8: Model Performance

- Input probability space: $\mathbb{X}_8 \subseteq \mathbb{R}^m$ which includes the test dataset from the output of ML_1, $X_{test} \in \mathbb{Y}_1$, and $X_{train\ trend} \in \mathbb{Y}_6$, which is the output of ML_6 as the trend prediction of the test monthly sea level data.
- Output probability space: $\mathbb{Y}_8 \subseteq \mathbb{R}^m$ which is the time series of the best model.
- Learning morphism: F_{ML_8}: $X_{test\ trend} = X_{test} - S_{avg} - R_{real}$

  The best $X_{train\ trend}$ is selected to be the output Y.
- Prior distribution on parameters: S_{avg} is a time series, calculated by averaging the monthly seasonal component from the training dataset; R_{real} is the time series representing the real residual component, obtained using ML_5 by differencing its output and input $X_5 - Y_5$.
- Empirical risk function: MAPE $\frac{1}{n}\Sigma\frac{|x_{test\ trend} - x_{train\ trend}|}{|x_{test\ trend}|} \times 100\%$, when MAPE is smaller, the performance of the model is better, and the model is chosen.

ML_9: Linear Regression

- Input probability space: $\mathbb{X}_9 \in \mathbb{Y}_7 \subseteq \mathbb{R}^m$ which is the predicted time series trend as the output of ML_7
- Output probability space: $\mathbb{Y}_9 \subseteq \mathbb{R}$ which is a value predicting the annually increase in sea level

- Learning morphism: $\widehat{y_t} = kx_t + b$, where $x \in \mathbb{X}_9$, $y = k \times 12 \in \mathbb{Y}_9$
- Prior distribution on parameters: k and b are real values $\in \mathbb{R}$
- Empirical risk function: MAPE $\frac{1}{n}\Sigma\frac{|x_t - \widehat{y_t}|}{|x_t|} \times 100\%$

MLM: Workflow

- Input probability space: $\mathbb{X}_0 \subseteq \mathbb{R}^m$, it has three sets of time series data (monthly sea level data, monthly air temperature data, and monthly water temperature data), and includes the raw training dataset and the raw testing dataset.
- Output probability space: $\mathbb{Y}_9 \subseteq \mathbb{R}$, it is a prediction for the sea level elevation
- Use ML_0 for data cleaning, deleting the data points with missing values and the data points being severely influenced by El Niño events.
- Use ML_1 (Kalman Filter) to fill in the blank data points deleted by ML_0 for all three datasets.
- Use ML_2 (ADF test) to check for the stationary of the output of ML_1 and all datasets are stationary.
- Use ML_4 (STL decomposition) to decompose the training datasets from the outputs of ML_1 into trend, seasonal, and residual components utilizing ML_3 (LOESS Model).
- Since in this project, the ACF from ML_4 shows auto-correlation in the residual parts for all three datasets, use ML_2 to ensure the stationarity of the residuals, then use ML_5 (ARMA model) to fit the residual parts to separate the auto-aorrelation part from the white noise in the residual; the white noise is later used as the real residual component.
- Use ML_6 (ARIMAX model) to fit three models:
  - the trend of monthly air temperature as a function of the trend of monthly sea level ($T_s \sim T_a$);
  - the trend of monthly water temperature as a function of the trend of monthly sea level ($T_s \sim T_w$);
  - the trend of the combination of the monthly air and water temperatures as a function of the trend of monthly sea level ($T_s \sim T_a + T_w$);
  - Use ML_7 (Forward rolling-window) to cross-validate the best model.
- Use ML_8 for the test, the second model $T_s \sim T_w$ performs the best and is selected.
- Use ML_9 (Linear Regression) to obtain the predicted sea level rise for the next year.

Therefore, the workflow of the MLM models in this project is
$ML\colon \mathbb{X}_0 \rightarrow \mathbb{Y}_9 = ML_9 \circ ML_8 \circ ML_7 \circ ML_6 \circ ML_5 \circ ML_4 \circ ML_3 \circ ML_2 \circ ML_1 \circ ML_0$

Besides, as an acknowledgment to the MLM section, most of the MLM models have references. The MLM for Kalman filter, ARMA and ARIMAX models are written according to Carmona (2014). The MLM for ADF test is written according to Said and Dickey (1984). The MLM for the LOESS is written according to NIST (2017). The MLM for the STL decomposition is written according to Ouyang (2021).

Result and Insights

Model performance

After finding the best ARIMAX models for the three relationships ($T_s(t) \sim T_w(t)$, $T_s(t) \sim T_a(t)$, and $T_s(t) \sim T_w(t)\ +\ T_a(t)$) using cross-validation, the performances of these models are evaluated using MAPE and RMSE and are shown in Table 3. Notably, the evaluations are based on the test dataset, and hence the test data is also processed into trends since the output of the ARIMAX models are trends.

The seasonal components of sea level, air temperature, and water temperature from STL decomposition using the training dataset are evolved into the average seasonal component of sea level S_s, the average seasonal component of air temperature S_a, and the average seasonal component of water temperature S_w. Particularly, average seasonal components are averaged for each month, and the results are time series that describes the seasonal components for 12 months from January to December.

Furthermore, the residual for each dataset is also obtained. In the STL: Remainder section, the residuals from STL decomposition are further processed using ARMA model, and the residuals (from the ARMA model) of the residuals (from the STL decomposition model) are white noises as a result. The white noise from the ARMA model is therefore the real residual since the original residual from the STL model contains autocorrelations that could be trend or seasonal components and the ARMA model removes the part of data containing the autocorrelations. The white noise of sea level is R_s, water temperature is R_w, and air temperature is R_a. The test datasets are then converted into trends by subtracting the averaged seasonal components and the real residuals. With the test dataset being X_{test}, $t$he equations for the process is $T_s = X_{sea\ test} - R_s - S_s$, $T_w = X_{water\ test} - R_w - S_w$, and $T_a = X_{air\ test} - R_a - S_a$. This process eliminates the seasonal and residual components from the test datasets to transform them into trends, and is described as ML_8: model performance in the previous MLM section. The models are then evaluated using the processed test datasets. Results of this evaluation are shown in the Model Performance section in Table 3.

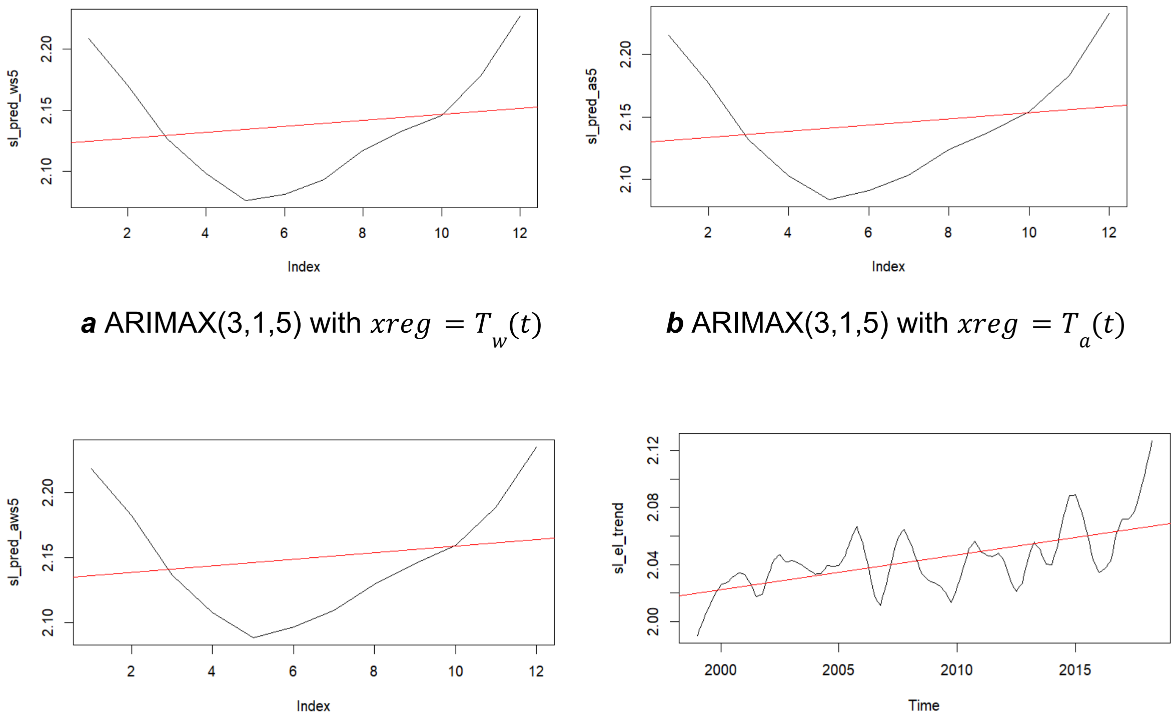
Table 3. The training and test performance of the ARIMAX model

Method	Model		MAPE	RMSE
ARIMAX(3, 1, 5)	$T_s(t) \sim T_w(t)$	Training	0.01319473	0.03872349
		Test	0.01722064	0.04714598
ARIMAX(3, 1, 5)	$T_s(t) \sim T_a(t)$	Training	0.01319571	0.03873266
		Test	0.01720734	0.04721692
ARIMAX(3, 1, 5)	$T_s(t) \sim T_w(t) + T_a(t)$	Training	0.01594713	0.05172952
		Test	0.01773	0.04832676

According to Table 3, the test dataset shows the best results in the ARIMAX(3,1,5) model when $T_s(t) \sim T_w(t)$ while the other two models exhibit similar performances. Final prediction is therefore made based on the $T_s(t) \sim T_w(t)$ model, and the predictions from other models are also made and shown in this project. Remarkably, all the models yield small MAPEs and RMSEs with them being close to 0, indicating the models accurately capture the underlying patterns of the increase in Tuvalu sea level according to the increase in temperatures. The predictions made based on these models are then likely to be accurate.

Predictions

The predictions made by the models are then plotted in Figure 14. The prediction is an annual prediction that lasts from March, 2023 to February, 2024 with the dataset used in this project ending in February, 2023. Linear regression is utilized in the predictions to calculate the increasing slope of the sea level.



a ARIMAX(3,1,5) with $xreg = T_w(t)$ ***b*** ARIMAX(3,1,5) with $xreg = T_a(t)$

c ARIMAX(3,1,5) with $xreg = T_w(t) + T_a(t)$ ***d*** STL+Random Walk

Figure 14. One year prediction by STL decomposition and ARIMAX models.

The dotted black line in Figure 14 represents the predicted sea level trend, and the dotted red line in Figure 14 represents the linear regression of the trend lines. Figure 14a-c are predictions generated by the ARIMAX models (T_s as a function of T_w, T_a, and their combination) and the plots include only the predicted year, and Figure 14d is the prediction made by STL decomposition and the plot has the predicted year and the past years. It is clear that the predictions produced by the ARIMAX models are close to the sea level trend in reality, and therefore the predictions based on the ARIMAX models are better than the prediction based on the STL decomposition model. From the linear regression of the trends, it is evident that sea level slightly increases in all models. The annual increases are calculated numerically based on the linear regressions and are listed in Table 4.

Table 4. The predictive results of the STL+Random Walk, the STL+ARIMAX model

Method	Model	Annual Increase (m)
ARIMAX(3, 1, 5)	$T_s(t) \sim T_w(t)$	0.02980674
ARIMAX(3, 1, 5)	$T_s(t) \sim T_a(t)$	0.02961462
ARIMAX(3, 1, 5)	$T_s(t) \sim T_w(t)\ +\ T_a(t)$	0.03043115
STL Decomposition	$T_s(t)$	0.02927214

From the table, the sea level increase for all 4 models are approximately 3cm each year. Since the previous results suggest that the $T_s(t) \sim T_w(t)$ model is the best, the final result of this model is that the Tuvalu sea level increases 2.98cm per year. Remarkably, the sea level of Tuvalu ranges from 1.8m to 2.3m.

Conclusions

Issue raised

Although the $T_s(t) \sim T_w(t)$ ARIMAX model is determined to be the best model based on its performance, the predictions are similar as indicated in Table 4. The main reason for the little difference lies in the model selection. The cross-validation results suggest the best $p,\ d,$ and q are 3, 1, and 5 for all ARIMAX models. Since the long-term trend of temperature change is caused by global warming, it is expected that the air temperature and water temperatures share a similar trend in the long term. Therefore, in terms of taking the sea level as a function of the air or water temperature, the functions ought to be similar. Despite the reasoning, more models can still be developed to investigate the relationships as an improvement of this project.

Another issue raised in this project is that the prediction made based on the models is for now a one-year prediction, while the goal of the project is to predict the long-term sea level change in accordance with the change in temperature caused by global warming. The model designed in this project can be further improved to make direct predictions that last for decades or hundreds of years to accurately predict the sinking of Tuvalu straightaway. Currently, based on the result that the sea level raises 2.98cm per year, and the fact that the highest point in Tuvalu raises 4.6m above the sea level, the conclusion is that Tuvalu is predicted to sink entirely beneath the sea after 154 years.

Lastly, the project only concerns the influence of global warming on the change in Tuvalu sea level. Since one of the main purposes of the project is to develop a model that can be used as an early warning system, the sinking time prediction should be as precise as possible to help the Tuvalu government, Tuvalu residents, and the international societies prepare for the sinking. Although global warming is the main cause of sea level rise, as mentioned in the Introduction section, sea level is also influenced by daily tides, weather, vertical land movements, and so on, and these factors are not considered in this project on purpose. In terms of the accuracy of predictions, the model can be promoted by incorporating these features as factors into the model.

Importance

Regardless of the issues mentioned above, from the Results section, the model designed by this project is still an accurate fit for the past observations and is capable of making accurate predictions for the future Tuvalu sea level. Employing the $T_s(t) \sim T_w(t)$ STL+ARIMAX(3,1,5) model, the purpose of the project is fulfilled, the decisions mentioned in the Introduction section will be impacted and the business values will be influenced. By accurately depicting the trend of Tuvalu sea level, this project provides the timeline for the resettlement of Tuvalu residents and the adaptations the government should do to slow down the sinking, and also provides references for other sinking island countries. In addition to the impacts, this project presents the severity of global warming by demonstrating its effects on Tuvalu, which will result in the submersion of the country and the displacement of over ten thousand individuals from their homes in the foreseeable future.

Besides, this project provides insights into the STL decomposition model and the ARIMAX model. This project represents a significant advancement in predicting the rise of Tuvalu sea level, as it utilizes the two models for the first time. The project eliminates the seasonal influence on sea level changes to make accurate predictions according to the trends. Furthermore, the STL model is robust to outliers, while outliers are inevitable in real life. The model's robustness against outliers enhances its ability to identify the underlying pattern of the trend of sea level, thereby, improving the accuracy of the prediction. In conclusion, applying these two models has significantly strengthened the accuracy of predicting the Tuvalu sea level changes. This project can be served as a reference to make sea level predictions for the following research in relevant fields.

Moreover, this project is the first one to forecast the rise of Tuvalu sea level based on temperature changes caused by global warming. As it is known, global warming is the main cause of the rise of Tuvalu sea level. Depicting the change in sea level according to the temperature change is the most accurate way when making predictions for the sea level changes in the long term. This point can also be considered as a reference for other research when analyzing the rise of sea levels to make accurate forecasts.

In summary, this project successfully employs a combination of the STL decomposition model and the ARIMAX model to predict the future sea level of Tuvalu with high accuracy, taking into account the changes in air and water temperature. This approach provides a comprehensive understanding of the relationship between global warming and sea level rise, and highlights the importance of incorporating temperature data in such predictions. Ultimately, this project has significant implications for the policymakers in Tuvalu as well as in other countries around the world, for it emphasizes the urgent need for actions to mitigate the effects of climate change and protect vulnerable regions like Tuvalu.

Acknowledgment

The code used for data preprocessing, implementation of the models, data analysis, and forecasts is available on GitHub at Huanyue17/ESE527-Tuvalu-Sea-Level: This repository is for ESE527. (github.com). The data used in this study are publicly available at Monthly sea levels for TUVALU (bom.gov.au).

We would like to express our most sincere gratitude to Dr. Patricio La Rosa for providing invaluable suggestions and insightful guidance for this project. We are also grateful to Dr. Patricio La Rosa and Dr. Balathasan Giritharan for the informative lectures they provided, which greatly contributed to our understanding of the techniques used in the project.

References

Aung, Than, et al. “Sea Level Threat in Tuvalu.” *American Journal of Applied Sciences*, vol. 6, no. 6, June 2009, pp. 1169–74, doi:https://doi.org/10.3844/ajassp.2009.1169.1174.

Caldwell, Patrick C.; Merrifield, Mark A.; Thompson, Philip R. 2001. Sea level measured by tide gauges from global oceans as part of the Joint Archive for Sea Level (JASL) since 1846. *NOAA National Centers for Environmental Information*. Dataset. https://doi.org/10.7289/v5v40s7w.

Carmona, René. Statistical Analysis of Financial Data in R. *Springer Texts in Statistics, New York, NY, Springer New York*, 2014. https://link.springer.com/book/10.1007/978-1-4614-8788-3

Cawi, Eric, et al. “Designing Machine Learning Workflows with an Application to Topological Data Analysis.” PLOS ONE, vol. 14, no. 12, 2 Dec. 2019, p. e0225577, https://doi.org/10.1371/journal.pone.0225577

Cooley, Sarah, et al. "Oceans and coastal ecosystems and their services." *IPCC AR6 WGII.* Cambridge University Press, 2022, p. 393. https://awi.eprints-hosting.org/id/eprint/56137/1/IPCC_AR6_WGII_Chapter03.pdf

Current and Future Climate of Tuvalu. 4 Jan. 2014, https://world.350.org/pacific/files/2014/01/4_PCCSP_Tuvalu_8pp.pdf.

Fisheries Department | Tuvalu Fisheries. https://tuvalufisheries.tv/tag/fisheries-department/.

G, Vijay Kumar. “Stationarity: Statistical Tests to Check Stationarity in Time Series.” *Analytics Vidhya*, 14 Mar. 2023, https://www.analyticsvidhya.com/blog/2021/06/statistical-tests-to-check-stationarity-in-time-series-part-1/.

Gallagher, Sean. “Seas Rise, Hope Sinks: Tuvalu’s Vanishing Islands – in Pictures.” *The Guardian,* 27 May 2019, www.theguardian.com/global-development/gallery/2019/may/27/seas-rise-hope-sinks-tuvalu-vanishing-islands-in-pictures.

Hunter, John., “A note on Relative Sea Level Change at Funafuti, Tuvalu.” *Antarctic Cooperative Research Centre. University of Tasmania*, 2002, p. 25. http://staff.acecrc.org.au/~johunter/tuvalu.pdf

Hyndman, R.J., & Athanasopoulos, G. *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. 2018. OTexts.com/fpp2.

Hyndman, R. J., & Athanasopoulos, G. 2021. *Forecasting: Principles and practice*. Amazon. https://www.amazon.com/Forecasting-Principles-Practice-Rob-Hyndman/dp/0987507133

Keith, Michael. "Model Validation Techniques for Time Series." *Medium*, Towards Data Science, 27 June 2022, https://towardsdatascience.com/model-validation-techniques-for-time-series-3518269bd5b3.

Lawler, Gregory F., and Vlada Limic. *Random Walk: A Modern Introduction.* Google Books, Cambridge University Press, 24 June 2010, https://books.google.com/books?id=UBQdwAZDeOEC&lpg=PR5&ots=Qg0DPjRGYp&dq=random%20walk&lr&pg=PA2#v=onepage&q&f=false.

Mushtaq, Rizwan. "Augmented Dickey Fuller Test." *SSRN*, 17 Aug. 2011, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1911068.

NIST, "LOESS (aka LOWESS)", section 4.1.4.4, *NIST/SEMATECH e-Handbook of Statistical Methods*, accessed 14 April 2017, https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm.

NOAA Fisheries. "National Fishery Sector Overview Tuvalu| NOAA Fisheries." *Noaa.gov*, Jan. 2010, www.fisheries.noaa.gov/about-us.

Riebeek, Holli. "Global Warming." *Nasa*, NASA Earth Observatory, 3 June 2010, https://earthobservatory.nasa.gov/features/GlobalWarming

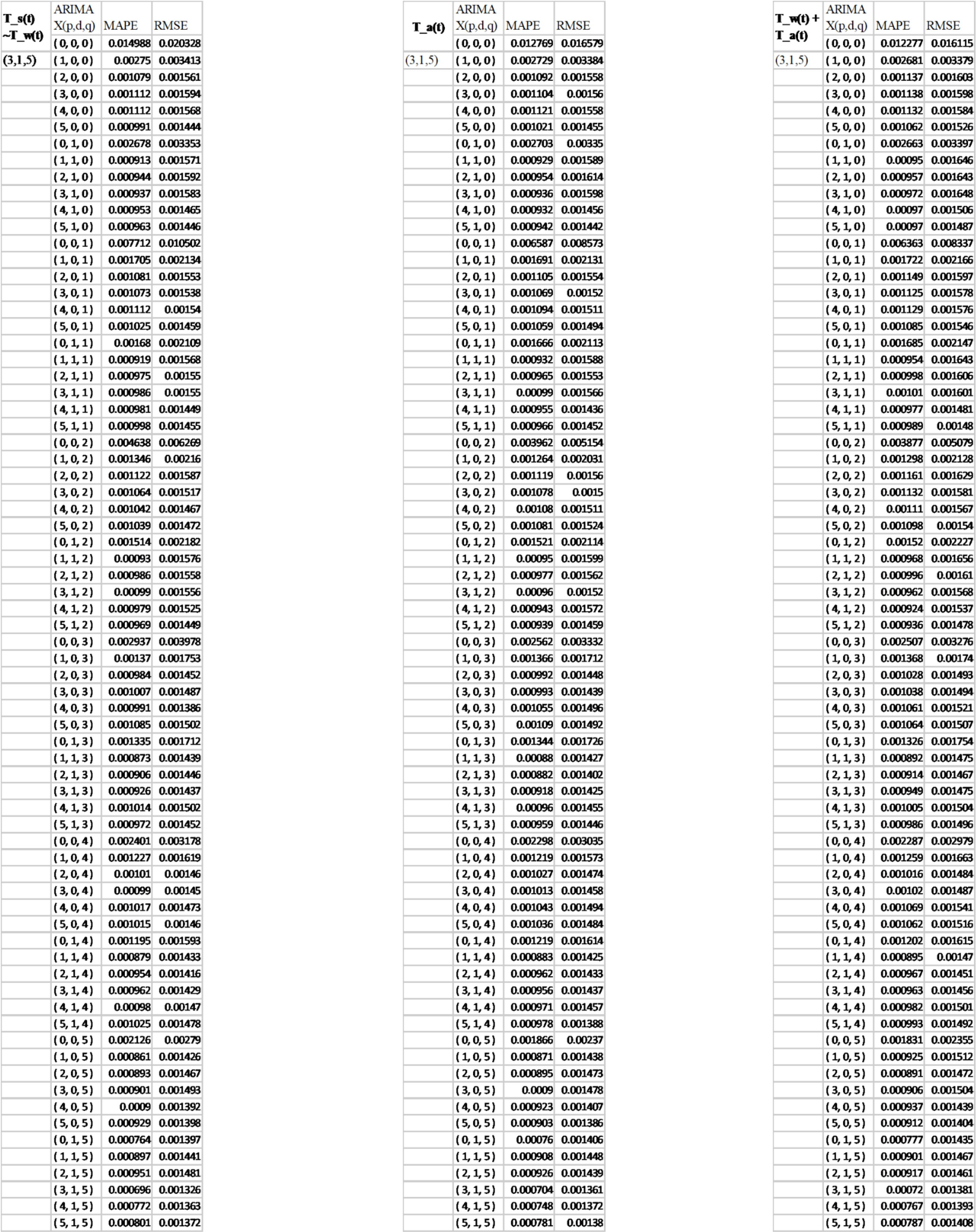
Said, E., and David A. Dickey. "Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order." *Biometrika*, vol. 71, no. 3, Dec. 1984, p. 599, https://doi.org/10.2307/2336570.

Singh, Awnesh, and Than Aung. "Effect of Barometric Pressure on Sea Level Variations in the Pacific Region." *The South Pacific Journal of Natural and Applied Sciences*, vol. 23, no. 1, 2005, p. 9, doi:https://doi.org/10.1071/sp05002.

Ouyang, Zuokun, et al. "STL Decomposition of Time Series Can Benefit Forecasting Done by Statistical Methods but Not by Machine Learning Ones." *MDPI, Multidisciplinary Digital Publishing Institute*, 8 July 2021, https://www.mdpi.com/2673-4591/5/1/42.

Zvornicanin, Written by: Enes. "Choosing the Best Q and P from ACF and PACF Plots in Arma-Type Modeling." *Baeldung on Computer Science*, 8 Nov. 2022, www.baeldung.com/cs/acf-pacf-plots-arma-modeling.

Appendix A: Results of Cross-Validation



T_s(t) ~T_w(t)	ARIMA X(p,d,q)	MAPE	RMSE
	(0, 0, 0)	0.014988	0.020328
(3,1,5)	(1, 0, 0)	0.00275	0.003413
	(2, 0, 0)	0.001079	0.001561
	(3, 0, 0)	0.001112	0.001594
	(4, 0, 0)	0.001112	0.001568
	(5, 0, 0)	0.000991	0.001444
	(0, 1, 0)	0.002678	0.003353
	(1, 1, 0)	0.000913	0.001571
	(2, 1, 0)	0.000944	0.001592
	(3, 1, 0)	0.000937	0.001583
	(4, 1, 0)	0.000953	0.001465
	(5, 1, 0)	0.000963	0.001446
	(0, 0, 1)	0.007712	0.010502
	(1, 0, 1)	0.001705	0.002134
	(2, 0, 1)	0.001081	0.001553
	(3, 0, 1)	0.001073	0.001538
	(4, 0, 1)	0.001112	0.00154
	(5, 0, 1)	0.001025	0.001459
	(0, 1, 1)	0.00168	0.002109
	(1, 1, 1)	0.000919	0.001568
	(2, 1, 1)	0.000975	0.00155
	(3, 1, 1)	0.000986	0.00155
	(4, 1, 1)	0.000981	0.001449
	(5, 1, 1)	0.000998	0.001455
	(0, 0, 2)	0.004638	0.006269
	(1, 0, 2)	0.001346	0.00216
	(2, 0, 2)	0.001122	0.001587
	(3, 0, 2)	0.001064	0.001517
	(4, 0, 2)	0.001042	0.001467
	(5, 0, 2)	0.001039	0.001472
	(0, 1, 2)	0.001514	0.002182
	(1, 1, 2)	0.00093	0.001576
	(2, 1, 2)	0.000986	0.001558
	(3, 1, 2)	0.00099	0.001556
	(4, 1, 2)	0.000979	0.001525
	(5, 1, 2)	0.000969	0.001449
	(0, 0, 3)	0.002937	0.003978
	(1, 0, 3)	0.00137	0.001753
	(2, 0, 3)	0.000984	0.001452
	(3, 0, 3)	0.001007	0.001487
	(4, 0, 3)	0.000991	0.001386
	(5, 0, 3)	0.001085	0.001502
	(0, 1, 3)	0.001335	0.001712
	(1, 1, 3)	0.000873	0.001439
	(2, 1, 3)	0.000906	0.001446
	(3, 1, 3)	0.000926	0.001437
	(4, 1, 3)	0.001014	0.001502
	(5, 1, 3)	0.000972	0.001452
	(0, 0, 4)	0.002401	0.003178
	(1, 0, 4)	0.001227	0.001619
	(2, 0, 4)	0.00101	0.00146
	(3, 0, 4)	0.00099	0.00145
	(4, 0, 4)	0.001017	0.001473
	(5, 0, 4)	0.001015	0.00146
	(0, 1, 4)	0.001195	0.001593
	(1, 1, 4)	0.000879	0.001433
	(2, 1, 4)	0.000954	0.001416
	(3, 1, 4)	0.000962	0.001429
	(4, 1, 4)	0.00098	0.00147
	(5, 1, 4)	0.001025	0.001478
	(0, 0, 5)	0.002126	0.00279
	(1, 0, 5)	0.000861	0.001426
	(2, 0, 5)	0.000893	0.001467
	(3, 0, 5)	0.000901	0.001493
	(4, 0, 5)	0.0009	0.001392
	(5, 0, 5)	0.000929	0.001398
	(0, 1, 5)	0.000764	0.001397
	(1, 1, 5)	0.000897	0.001441
	(2, 1, 5)	0.000951	0.001481
	(3, 1, 5)	0.000696	0.001326
	(4, 1, 5)	0.000772	0.001363
	(5, 1, 5)	0.000801	0.001372

T_a(t)	ARIMA X(p,d,q)	MAPE	RMSE
	(0, 0, 0)	0.012769	0.016579
(3,1,5)	(1, 0, 0)	0.002729	0.003384
	(2, 0, 0)	0.001092	0.001558
	(3, 0, 0)	0.001104	0.00156
	(4, 0, 0)	0.001121	0.001558
	(5, 0, 0)	0.001021	0.001455
	(0, 1, 0)	0.002703	0.00335
	(1, 1, 0)	0.000929	0.001589
	(2, 1, 0)	0.000954	0.001614
	(3, 1, 0)	0.000936	0.001598
	(4, 1, 0)	0.000932	0.001456
	(5, 1, 0)	0.000942	0.001442
	(0, 0, 1)	0.006587	0.008573
	(1, 0, 1)	0.001691	0.002131
	(2, 0, 1)	0.001105	0.001554
	(3, 0, 1)	0.001069	0.00152
	(4, 0, 1)	0.001094	0.001511
	(5, 0, 1)	0.001059	0.001494
	(0, 1, 1)	0.001666	0.002113
	(1, 1, 1)	0.000932	0.001588
	(2, 1, 1)	0.000965	0.001553
	(3, 1, 1)	0.00099	0.001566
	(4, 1, 1)	0.000955	0.001436
	(5, 1, 1)	0.000966	0.001452
	(0, 0, 2)	0.003962	0.005154
	(1, 0, 2)	0.001264	0.002031
	(2, 0, 2)	0.001119	0.00156
	(3, 0, 2)	0.001078	0.0015
	(4, 0, 2)	0.00108	0.001511
	(5, 0, 2)	0.001081	0.001524
	(0, 1, 2)	0.001521	0.002114
	(1, 1, 2)	0.00095	0.001599
	(2, 1, 2)	0.000977	0.001562
	(3, 1, 2)	0.00096	0.00152
	(4, 1, 2)	0.000943	0.001572
	(5, 1, 2)	0.000939	0.001459
	(0, 0, 3)	0.002562	0.003332
	(1, 0, 3)	0.001366	0.001712
	(2, 0, 3)	0.000992	0.001448
	(3, 0, 3)	0.000993	0.001439
	(4, 0, 3)	0.001055	0.001496
	(5, 0, 3)	0.00109	0.001492
	(0, 1, 3)	0.001344	0.001726
	(1, 1, 3)	0.00088	0.001427
	(2, 1, 3)	0.000882	0.001402
	(3, 1, 3)	0.000918	0.001425
	(4, 1, 3)	0.00096	0.001455
	(5, 1, 3)	0.000959	0.001446
	(0, 0, 4)	0.002298	0.003035
	(1, 0, 4)	0.001219	0.001573
	(2, 0, 4)	0.001027	0.001474
	(3, 0, 4)	0.001013	0.001458
	(4, 0, 4)	0.001043	0.001494
	(5, 0, 4)	0.001036	0.001484
	(0, 1, 4)	0.001219	0.001614
	(1, 1, 4)	0.000883	0.001425
	(2, 1, 4)	0.000962	0.001433
	(3, 1, 4)	0.000956	0.001437
	(4, 1, 4)	0.000971	0.001457
	(5, 1, 4)	0.000978	0.001388
	(0, 0, 5)	0.001866	0.00237
	(1, 0, 5)	0.000871	0.001438
	(2, 0, 5)	0.000895	0.001473
	(3, 0, 5)	0.0009	0.001478
	(4, 0, 5)	0.000923	0.001407
	(5, 0, 5)	0.000903	0.001386
	(0, 1, 5)	0.00076	0.001406
	(1, 1, 5)	0.000908	0.001448
	(2, 1, 5)	0.000926	0.001439
	(3, 1, 5)	0.000704	0.001361
	(4, 1, 5)	0.000748	0.001372
	(5, 1, 5)	0.000781	0.00138

T_w(t) + T_a(t)	ARIMA X(p,d,q)	MAPE	RMSE
	(0, 0, 0)	0.012277	0.016115
(3,1,5)	(1, 0, 0)	0.002681	0.003379
	(2, 0, 0)	0.001137	0.001603
	(3, 0, 0)	0.001138	0.001598
	(4, 0, 0)	0.001132	0.001584
	(5, 0, 0)	0.001062	0.001526
	(0, 1, 0)	0.002663	0.003397
	(1, 1, 0)	0.00095	0.001646
	(2, 1, 0)	0.000957	0.001643
	(3, 1, 0)	0.000972	0.001648
	(4, 1, 0)	0.00097	0.001506
	(5, 1, 0)	0.00097	0.001487
	(0, 0, 1)	0.006363	0.008337
	(1, 0, 1)	0.001722	0.002166
	(2, 0, 1)	0.001149	0.001597
	(3, 0, 1)	0.001125	0.001578
	(4, 0, 1)	0.001129	0.001576
	(5, 0, 1)	0.001085	0.001546
	(0, 1, 1)	0.001685	0.002147
	(1, 1, 1)	0.000954	0.001643
	(2, 1, 1)	0.000998	0.001606
	(3, 1, 1)	0.00101	0.001601
	(4, 1, 1)	0.000977	0.001481
	(5, 1, 1)	0.000989	0.00148
	(0, 0, 2)	0.003877	0.005079
	(1, 0, 2)	0.001298	0.002128
	(2, 0, 2)	0.001161	0.001629
	(3, 0, 2)	0.001132	0.001581
	(4, 0, 2)	0.00111	0.001567
	(5, 0, 2)	0.001098	0.00154
	(0, 1, 2)	0.00152	0.002227
	(1, 1, 2)	0.000968	0.001656
	(2, 1, 2)	0.000996	0.00161
	(3, 1, 2)	0.000962	0.001568
	(4, 1, 2)	0.000924	0.001537
	(5, 1, 2)	0.000936	0.001478
	(0, 0, 3)	0.002507	0.003276
	(1, 0, 3)	0.001368	0.00174
	(2, 0, 3)	0.001028	0.001493
	(3, 0, 3)	0.001038	0.001494
	(4, 0, 3)	0.001061	0.001521
	(5, 0, 3)	0.001064	0.001507
	(0, 1, 3)	0.001326	0.001754
	(1, 1, 3)	0.000892	0.001475
	(2, 1, 3)	0.000914	0.001467
	(3, 1, 3)	0.000949	0.001475
	(4, 1, 3)	0.001005	0.001504
	(5, 1, 3)	0.000986	0.001496
	(0, 0, 4)	0.002287	0.002979
	(1, 0, 4)	0.001259	0.001663
	(2, 0, 4)	0.001016	0.001484
	(3, 0, 4)	0.00102	0.001487
	(4, 0, 4)	0.001069	0.001541
	(5, 0, 4)	0.001062	0.001516
	(0, 1, 4)	0.001202	0.001615
	(1, 1, 4)	0.000895	0.00147
	(2, 1, 4)	0.000967	0.001451
	(3, 1, 4)	0.000963	0.001456
	(4, 1, 4)	0.000982	0.001501
	(5, 1, 4)	0.000993	0.001492
	(0, 0, 5)	0.001831	0.002355
	(1, 0, 5)	0.000925	0.001512
	(2, 0, 5)	0.000891	0.001472
	(3, 0, 5)	0.000906	0.001504
	(4, 0, 5)	0.000937	0.001439
	(5, 0, 5)	0.000912	0.001404
	(0, 1, 5)	0.000777	0.001435
	(1, 1, 5)	0.000901	0.001467
	(2, 1, 5)	0.000917	0.001461
	(3, 1, 5)	0.00072	0.001381
	(4, 1, 5)	0.000767	0.001393
	(5, 1, 5)	0.000787	0.001408

Figure I. The detailed cross-validation results for each model ($T_s \sim T_w$, $T_s \sim T_a$, $T_s \sim T_a + T_w$)