Abstract—A slight change in sea surface temperature (SST) is a critical condition as too high of SST could cause coral bleaching, which could result in the declining numbers of fish individuals and species on coral reefs. The people of Kiribati, one of the most affected countries by climate change, depend on the reef and ocean for food and economics, thus advance knowledge of SST in the region could benefit the country. In this paper, a multiple linear regression (MLR) is developed for forecasting the sea surface temperature anomaly (SSTA) of the Kiribati Region (7° N-15° S, 150° W-170° E) using the Sea Level Pressure Anomaly (SLPA), Air Temperature Anomaly (ATA), Total Cloudiness Anomaly (TCA), Relative Humidity Anomaly (RHA), Wind Eastward component Anomaly (WECA), Wind Northward component Anomaly (WNCA) and Wind Scalar Anomaly (WSA) as predictors. We validate the proposed model and determine which predictors should we include, and to what extend does this model predict the SSTA in the Kiribati region. The proposed model is compared with the Naïve Method by various error functions such as the Root Square Mean Error (RSME), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). We found that ATA, TCA, WNCA, WECA and WSA are the best predictor variables in the forecasting model, which satisfy all the MLR assumptions and performing better than the Naïve method, hence, it may be useful for forecasting SST accurately.

Keywords—Modeling Sea surface temperature, Multiple linear regression, Naïve Method, Error functions.

I. INTRODUCTION

Accurate predictions of upper ocean conditions in the Kiribati ocean, including sea surface temperature anomalies (SSTA), defined here as departures from monthly climatology, could provide many economic and societal benefits for Kiribati. For example, an I-Kiribati life is depending on the reef and ocean, so forecasts based on the state of the climate system, especially the state of El Niño–Southern Oscillation (ENSO) (Trenberth, 1997) and SSTs in general have the potential to provide useful information for fishermen, water resource managers and industry (Bell et al., 2011). Coral bleaching in Kiribati remains at Alert level 2 in the Line and Phoenix islands in Kiribati but since coral bleaching is associated with SST, the future knowledge of SST will alert the government to make appropriate decision for the people (Keith et al., 2018).

Since early 1980s, several studies have been conducted in developing forecasting models for SST anomalies to understand the ENSO phenomena and to monitor the effect of SST fluctuation in our oceans. According to the literature (Barnston & Ropelewski, 1992; Latif et al., 1994; Martinez et al., 2009) the models can be classified into 4 distinct classes: simple dynamical coupled model (Berliner et al., 2000); coupled GCMs (general circulation models) (Rosati et al., 1997); hybrid coupled models (Neelin, 1990; Barnett et al., 1993) and statistical models (Barnston et al., 1999).

Various GCMs models developed including those using atmospheric models forced with other observed ocean variables such as heat flux and wind; while some are using atmospheric models forced with SST (Rosati et al., 1997). Neelin (1992) review the status of a variety of coupled models used and Barnett (1993) explained that hybrid coupled model is more like an extended version of GCMs in which a complete ocean general circulation model (OGCM) is coupled with a simpler atmospheric model such as statistical models represented by correlations between SST and wind stress.

Barnston and Ropelewski (1992) presented the performance comparison between dynamical and statistical models and concluded that both gives comparable skills, however they do not predict the 1996-97 El Nino so more work is needed to improve the models. Moreno (2015) argued that Dynamical models are unable to properly reproduce tropical climate variability hence statistical methodologies emerge as an alternative to improve the SST predictability.

Other statistical models which has been stated in the literature involve the Markov approach (Xue et al., 2000) where the author uses autocorrelation and autocovariance in east Pacific SST, but wind stress and sea-level indices are also used as predictors. Furthermore, studies (e.g. Barnston et al., 1999) have developed a linear regression model using east equatorial Pacific wind stress and trend precursor tropical Pacific SST as predictors and Moreno (2015) introduced the S4CAST model which caters for the non-stationary links between the predictor and the predictant. In most of these studies, there has been very little efforts to forecast the SST in the Kiribati region, they either study the Pacific region as a whole or a slight smaller regions, but never to a local site.

In this paper, we study the SSTA variation in the Kiribati region focusing on developing an empirical statistical model, using MLR, to provide a baseline for studying SSTA in this region.
II. DATASET

The data is provided from International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 2.5 (R2.5), based on individual observations in the International Maritime Meteorological Archive (IMMA) format (see R2.5-imma). The data provided for the Kiribati region (7° N-15°S, 150° W-170° E), were the gridded 2° × 2° monthly observation values for the entire region (total of 220 grid points) for the eight parameters: Sea surface temperature (SST), Sea level pressure (SLP), Air temperature (AT), Total cloudiness (TC), Relative humidity (RH), Wind eastward component (WEC), Wind northward component (WNC) and Wind scalar (WS) for the years 1960 January to 2006 February. In this study we use the monthly anomaly data for each variable which was determined as follows:

\[ P_a = P - \bar{P} \]  
\[ \bar{P} = \frac{1}{n} \sum_{i=1}^{n} P_i \]

where \( P_a \) is the monthly anomaly reading of a parameter \( P \), \( \bar{P} \) is the climatology, average of \( P \) for the whole dataset period.

III. METHODOLOGY

Multiple linear regression (MLR) is a linear statistical technique that finds the best relationship between a dependent variable and several other independent or predictor variables based on the ordinary least square (OLS) method. The best fit in the OLS is obtained by minimizing the sum of the squared residuals, where a residual is the difference between an observed value and the fitted value from the regression. The MLR models can be presented by the following equation:

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i \]  

where the subscript \( i \) denotes the observational unit from which the observations on \( Y \) were taken. The second subscript designate the independent variable. The sample size is denoted with \( n \), \( i = 1, ..., n \), and \( p \) is the number of independent variables. In this study, we have in total 673 observational units with 7 independent variables.

A. Data Examination

Before we build a model, the data are examined to ensure that there are no missing values and that the distributions were appropriate.

B. Variable and Model selection

We constructed a regression model by identifying the independent variables to use, by the Pearson’s correlation between the variables and a backward selection method. The backward selection involve the significant statistical test of the coefficient of a variables \( \beta_0, \beta_1, ..., \beta_p \), where a variable is eliminated one at a time (the one with the highest \( p \)-value greater than 0.05 first) from the model until all the variable coefficients are statistically significant. In this paper, we used SPSS linear regression backward model procedure to determine the final model.

C. Examination of regression assumption

We examined that our model meet the assumptions for regression to justify the use for purposes of inference or prediction. Thus we need to show that the dependent variable \( Y \) and the \( p \) vector of independent variable \( X_j \) satisfied the following assumptions (Chaterjee et al., 2006; Rawlings et al., 2001):

1) Linearity: Multiple regression assumes that each predictor should show a linear relation with the outcome variable which can be examined by using scatterplots of the individual independent variable and the dependent variable.

2) Multi-collinearity: MLR assumes that there should be no strong correlation between the independent variables themselves. Failure to identify and report multicollinearity could result in misleading interpretations of the results (Mekanik, Imteaz, Gato-Trinidad, & Elmahdi, 2013). To check for multicollinearity, we inspect the correlation between the independent variables themselves and if the correlation value is more than 0.7 then there is a concern of collinearity. It is also confirmed by using the tolerance (T) and variance inflation factor (VIF) values given by the equations below:

\[ Tolerance = 1 - R^2, \]

\[ VIF = \frac{1}{Tolerance}. \]

where

\[ R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \]

\[ SSR = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \]

and

\[ SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2. \]

where \( Y_i \) is the actual value at case \( i \); \( \bar{Y} \) is the forecast value and \( \bar{Y} \) is the mean actual value with \( i = 1, 2, 3, ..., n \) for the forecast period. If the tolerance value is less than 0.10, or the VIF value is greater than 10, then there should be a concern.

3) Residual assumptions: This assumed that the errors, \( \epsilon_i \), are normally distributed with mean zero and constant variance, \( \sigma^2 \) which can be examined as follows:

a) Normality of errors: The errors, \( \epsilon_i \), at each set of values of the predictors, \( (x_{i1}, x_{i2}, \ldots) \), assumed to be normally distributed is examined by considering the PP plot, to see that our points (dots) lie reasonably close to the line of best fit that bisect the chart. We can also inspect the histogram of the error to confirm that the mean of residual is zero.
b) Homoscedasticity: It describes a situation in which the errors, $e_i$, at each set of values of the predictors, $(x_{1i}, x_{2i}, \ldots)$ have mean zero and constant variances. To check the homoscedasticity assumption, we inspect the residual scatter plot, and we would like to see a roughly rectangular distribution with most of data clustered in the center, without any pattern. Statistically, we confirm homoscedasticity by using Breusch-Pagan and the Koenker Tests. In this paper, we use the SPSS macro developed by Ahmand Daryanto (2018).

If any of these assumptions mentioned above are violated, then the forecasts, and scientific interpretation or conclusion drawn from the model may be considered inefficient, biased or misleading.

D. Further Model validation

We further examine the validity of our model to check how effective and reliable the model is, by examining the following:

1) Influential case: To determine whether the model fits the observed data well, or if it is influenced by a small number of cases, we examine the influential cases. Influential cases are extreme values which pull the regression line towards them therefore having a significant impact on the coefficients of the model. To identify these influential cases in this paper we are using the Cook’s distance which measures the difference between the regression coefficients obtained from the full data and the regression coefficients obtained by deleting the ith observation, or equivalently, the difference between the fitted values obtained from the full data and the fitted values obtained by deleting the ith observation.

2) Goodness of fit: We check how good our model generalizes, by examining $R^2$ and the Adj $R^2$. We need to observe that Adj $R^2$ value to be very close to $R^2$ (Field, 2013).

3) Test the significance of the model: The $F$-statistic is used to decide whether the model as a whole has statistically significant predictive capability. Under the null hypothesis that the model has no predictive capability, with $p$ numerator degrees of freedom and $n - p - 1$ denominator degrees of freedom. The null hypothesis is rejected if the $F$ ratio is large or when $p < 0.05$. We determined this by referring to the Analysis of variance (ANOVA) table of the model.

E. Comparison with Naïve Method

We compare our model with the Naïve method. Naïve forecasting is an estimating technique in which the last period’s actual values are used as this period’s forecast. The simplest Naïve Method, where the last actual datum is used to forecast the next period has an equation

$$\hat{Y}_{i+1} = Y_i$$  \hspace{1cm} (8)

To test the validity of the MLR SSTA model, the following error measurement statistics were considered:

1) Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$  \hspace{1cm} (9)

2) Mean Absolute Percentage Error:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{Y_i}$$  \hspace{1cm} (10)

3) Root Mean Square Deviation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$  \hspace{1cm} (11)

IV. RESULTS AND DISCUSSION

The multiple linear regression model was developed using the explanatory variables, the Sea level pressure anomaly (SLPA), Air temperature anomaly (ATA), Total cloudiness (TCA), Relative humidity anomaly, Wind eastward component anomaly (WECA), Wind northward component anomaly (WNCA) and Wind scalar anomaly (WSA) as our predictors with Sea surface temperature (SSTA) as our dependent variable (528x7=3696 data point). We use backward variable selection in SPSS to construct our final model to predict the SSTA in the Kiribati region.

A. Data Examination result

The descriptive statistics of all the variables presented in Table 1 were inspected. We found that our variables have no missing values. We are using the anomaly data for each variable hence we expect to have a mean of 0 for all with standard deviation ranges from 0.42 (ATA) to 1.49 (RHA). We noticed that not all the variables are normally distributed as shown with skewness magnitude ranges from 0.047 to 0.485, however this is not a problem since we are using a large sample.

<table>
<thead>
<tr>
<th>TABLE I: DESCRIPTIVE STATISTICS FOR THE ORIGINAL DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSTA</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>Valid</td>
</tr>
<tr>
<td>Missing</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Std. Deviation</td>
</tr>
<tr>
<td>Variance</td>
</tr>
<tr>
<td>Skewness</td>
</tr>
<tr>
<td>Kurtosis</td>
</tr>
<tr>
<td>Std. Error of Skewness</td>
</tr>
<tr>
<td>Std. Error of Kurtosis</td>
</tr>
</tbody>
</table>

B. Variable and Model selection result

By inspecting the Pearson’s correlation shown in Table 2, we found that ATA, TCA and WECA have correlation values more than 0.3. So these variables should be included in the model. Using SPSS backward selection procedure, the variables excluded (in order) are SLPA and RHA with p-values of 0.472 and 0.218 respectively. This implies that variables SLPA and RHA should not be used in the model, thus WNCA and WSA could also be used in our model.
By examining the values of the unstandardized coefficients for the models presented in Table 3, it implies that our final model is:

\[
SSTA = 1.025(ATA) + 0.101(TCA) \\
+ 0.056(\text{WECA}) + 0.049(\text{WNCA}) \\
+ 0.054(\text{WSA})
\] (12)

One could ask that since ATA, TCA, WECA were the only predictors having positive correlation values greater than 0.3 i.e., these variables are the only ones that correlate quite strongly with our dependent variable SSTA, why are we including WNCA and WSA? Even though WNCA and WSA have very small correlation with the dependent variable, they are still included in the model, because in combination with other variables they seems significantly correlate with SSTA, as shown in the Table 3.

### TABLE 3. REGRESSION COEFFICIENT FOR THE MODEL

<table>
<thead>
<tr>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>95.0% Confidence Interval for B</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>Lower Bound</td>
</tr>
<tr>
<td>(Constant)</td>
<td>-3.90E-10</td>
<td>0.008</td>
<td>0</td>
</tr>
<tr>
<td>ATA</td>
<td>1.025</td>
<td>0.022</td>
<td>0.857</td>
</tr>
<tr>
<td>TCA</td>
<td>0.101</td>
<td>0.027</td>
<td>0.086</td>
</tr>
<tr>
<td>WECA</td>
<td>0.056</td>
<td>0.012</td>
<td>0.122</td>
</tr>
<tr>
<td>WNCA</td>
<td>0.049</td>
<td>0.007</td>
<td>0.123</td>
</tr>
<tr>
<td>WSA</td>
<td>0.054</td>
<td>0.017</td>
<td>0.128</td>
</tr>
</tbody>
</table>

Dependent Variable: SSTA

### C. Examination of MLR assumptions

To draw conclusions about a population based on a regression analysis done on a sample, several assumptions must be true (Field, 2003; Rawlings et al., 2001):

1) **Linearity test**: We inspect linearity for each predictor separately by running scatterplots of each predictor with the dependent variable as shown in Fig. 1. It shows that our dependent variable is having a linear relationship with the independent variables. By inspecting the plot of residuals versus predicted as shown in Fig. 2, the plot is symmetrically distributed, with dots clustering in the middle of the plot without any pattern. This is a clear indication of randomness, which shows linearity between the dependent and independent variables.

![Fig 1. Scatterplots of each variable with the dependent variable](image-url)
be hidden by combining predictors into one variable, the predicted values.

![Scatterplots of Standardized residual vs Standardized Predicted value](image)

Fig. 2. Scatterplots of Standardized residual vs Standardized Predicted value

2) **Multi-collinearity test:** From the Pearson correlation as shown in Table 2, we could see that there is a slight correlation between some independent variables, for example TCA and SLPA has a correlation of 0.501. The correlation among the independent variables ranges in magnitude from 0.062 to 0.589 and none is more than 0.7, hence there might be no multi-collinearity among the independent variables which implies that we could use all the 7 independent variable as predictors in our model.

We further confirmed from the result of the collinearity statistics in Table 3. The VIF values ranges from 1.430 – 3.099, which are far above 0.1. The tolerance values ranges from 0.323 – 0.699 which are far less than 10, thus our model met the multi-collinearity assumption, i.e., we do not have multi-collinearity in our variables.

3) **Residual assumption tests:** We showed that the errors in our model can be considered normally distributed with mean of zero and constant variance $\sigma^2$.

   a) **Normality of error test:** By inspecting the histogram of the standardized residuals (Fig. 3) and the residual plot (Fig. 2) we can say that the error is normally distributed. The histogram shows that the distribution is normally distributed with mean of zero and constant standard deviation of 0.996. In the scatterplot of residual vs predicted value (Fig. 2), the distribution of the points is random and follow a rectangular shape clustering around zero.

   Furthermore, the PP plot (Fig. 4) confirmed the normality as the points (dots) lie reasonably close to the line of best fit that bisect the chart. There is no major deviation of the points from the line. This indicates that the residuals do not deviate much from normality, or we can say that it is normally distributed.

   b) **Homoscedasticity test:** From the scatterplot of the residuals (Fig. 3) there were no systematic pattern observed, which implies that our data met the homoscedasticity assumption.

   Using the Breusch-Pagan (BP) and the Koenker test (KT), we also confirmed that homoscedasticity can be observed in our data. The test used the Lagrange multiplier (LM), with null hypothesis assumed that heteroscedasticity not present and alternative hypothesis assumed heteroscedasticity. The LM gives a B-P value of 5.621 with significance of 0.345, and the KT value of 5.048 with significance of 0.410. Since both has significance more than 0.05, in both null hypothesis were not rejected.

   From all these results, we can conclude that our model met all the MLR assumptions hence could be considered reliable.

![The Histogram of the residuals](image)

Fig. 3. The Histogram of the residuals

![The P-P Plot of regression standardized residual](image)

Fig. 4. The P-P Plot of regression standardized residual

**D. Further Model validation**

We further check the validity of our model by examining the influential cases, goodness of fit and comparison with the Naive method.

1) **Influential cases:** In examining the same scatterplot Fig. 3, we can identify if there are any outliers that could influence our model. It was stated that outliers are cases that have a standardized residual lie outside a ±3.3 scale in the scatterplot (Statistics Solution., 2013). We observed that there are few points above 3.0.

   From Table 4, we have identified the three unusual cases with standardized residuals above 3.0, case number 102, 634 and 663, hence we need to verify whether these cases are influential case. By inspecting the Cook’s distance for the samples, if there are any with value more than one (>1), which
can be a problem, we observed that the maximum cook’s distance for all samples are far less than 1 as the maximum value is 0.029 shown in the Residual Statistics table, Table 5. This statistics implies that the three cases were not influential case to our model.

<table>
<thead>
<tr>
<th>Case Number</th>
<th>Std. Residual</th>
<th>SSTA</th>
<th>Predicted Value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>102</td>
<td>3.123</td>
<td>0.341</td>
<td>-0.2735</td>
<td>0.614891</td>
</tr>
<tr>
<td>634</td>
<td>3.403</td>
<td>0.811</td>
<td>0.14128</td>
<td>0.669895</td>
</tr>
<tr>
<td>663</td>
<td>3.176</td>
<td>1.131</td>
<td>0.50621</td>
<td>0.625218</td>
</tr>
</tbody>
</table>

Dependent Variable: SSTA

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Value</td>
<td>-1.68145</td>
<td>1.13953</td>
<td>0</td>
<td>465233</td>
</tr>
<tr>
<td>Residual</td>
<td>-0.552049</td>
<td>0.669895</td>
<td>0</td>
<td>196134</td>
</tr>
<tr>
<td>Deleted Residual</td>
<td>-0.557939</td>
<td>0.674261</td>
<td>7E-00</td>
<td>197862</td>
</tr>
<tr>
<td>Mahal. Distance</td>
<td>0.558</td>
<td>31.035</td>
<td>4.993</td>
<td>3.593</td>
</tr>
<tr>
<td>Cook’s Distance</td>
<td>0</td>
<td>0.029</td>
<td>0.001</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Dependent Variable: SSTA

2) Goodness of fit: The $R^2$ gives us a measure of how much of the variability in the outcome is accounted for by the predictors.

<table>
<thead>
<tr>
<th>R</th>
<th>R Square</th>
<th>Adj R Square</th>
<th>Std. Error</th>
<th>Durbin-Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>.921</td>
<td>.849</td>
<td>.848</td>
<td>.196868</td>
<td>1.044</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), WSA, ATA, WNCA, TCA, WECA

We examine how well our model generalizes by comparing the observed value of $R^2 = 0.849$ with the adjusted $\text{Adj} R^2 = 0.848$ which is very close indicating that the cross-validation of our model is quite good.

3) Model Significance test: From the analysis of variance results presented in Table 7, we found large value of $F$, which confirm that our model as a whole has statistically significant predictive capability.

<table>
<thead>
<tr>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>145.449</td>
<td>750.57</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>25.851</td>
<td>0.039</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>171.299</td>
<td>672</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent Variable: SSTA

a. Predictors: (Constant), WSA, ATA, WNCA, TCA, WECA

4) Comparison with Naïve Method: We also compare the performance of the proposed MLR model with a Naïve method as discussed in Section 3.6. The error measures RSME, MAE and MAPE presented in Table 8. The results show that the MLR model is performing better as it gives smaller error measures. We have shown that our MLR model is a valid model and in comparison against the Naïve method, we found that our proposed model perform better since it gives smaller values as shown in Table 8 below.

<table>
<thead>
<tr>
<th>RSME</th>
<th>MAE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLR</td>
<td>0.196</td>
<td>0.155</td>
</tr>
<tr>
<td>Naïve</td>
<td>0.227</td>
<td>0.182</td>
</tr>
</tbody>
</table>

The graph below shows the three models, the actual, the MLR and the naïve lines.

V. CONCLUSION

To model the SSTA in the Kiribati Ocean, in respect to other simultaneous variables, WSA, SLPA, RHA, ATA, WNCA, TCA, WECA, in this paper we proposed a multiple linear regression. A sample number of 673 observation for each variable were analyzed and used in the construction. The data for this analysis were obtained from International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 2.5 (R2.5).

This research found that the predictors ATA, TCA, WECA, WNCA, and WSA are the significant variables for forecasting SSTA in the Kiribati region. The proposed model with the sample data is found as shown below:

$$\text{SSTA} = 1.025(\text{ATA}) + 0.101(\text{TCA}) + 0.056(\text{WECA}) + 0.049(\text{WNCA}) + 0.054(\text{WSA})$$

We compare our model with Naïve model using the error measurement RMSE, MAE, and MAPE and found that the MLR model is performing well. We detected in our data several outliers, but we include them in our model since they are not influential cases, however it is interesting to know whether our model will be better if we do not include them, which could be considered in future studies.
REFERENCES


