Meteorological data prediction over selected stations in Sub-Saharan Africa: Leveraging on Machine Learning Algorithm

Segun Adebayo¹, Francis O. Aweda*², Isaac A. Ojedokun³ and James O. Agbolade⁴

¹Mechatronics Engineering Programme, College of Agriculture, Engineering and Science, Bowen University, Iwo, Nigeria
²Physics Programme, College of Agriculture, Engineering and Science, Bowen University, Iwo, Nigeria
³Electrical and Electronics Department, Federal University, Otuoke, Nigeria
⁴Electrical and Electronics Department, Federal Polytechnic, Ede, Osun State, Nigeria

*Correspondence: aweda.francis@bowen.edu.ng, https://orcid.org/0000-0003-3941-6647

Received: 24th January 2022, Revised: 15th November 2022, Accepted: 21st December 2022

Abstract: This study investigated selected meteorological data prediction leveraging on a Machine Learning Algorithm Approach over five selected stations in Nigeria. The algorithm of Machine Learning was explored using weather parameters such as temperature, wind speed, wind direction and relative humidity to predict the rainfall rate. In the results, five Gaussian models (i.e., Rational Quadratic, Squared Exponential, Matern 5/2, Exponential and Optimized GPR) revealed different Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE) with prediction speeds ranging from 15000 to 26000 and the training time included 7.936, 1.8923, 2.3701, 3.267 and 282.19, respectively. The predicted response as against the true response for the two models shows a linear graph passing through the origin which confirmed a perfect regression model, where all the points lie on a diagonal line. Therefore, the relationship between MSE, MAE and RMSE for different models revealed that the optimized GPR has a better performance as compared to others. More so, visualizing the relationship between the output variable (rainfall) and each input variable reveals that some input variables (relative humidity, rainfall, pressure, wind speed and direction) have a strong correlation with the output variable (rainfall), with others having a noisy relationship which is not very clear.

Keywords: Atmospheric Physics, Gaussian Model, Machine Learning, meteorological data, statistical model.

1 Introduction

Across the world, climate variability over spatial and temporal scale is usually detected from the analysis of long-term observational data of specific climatic variables over an averaging period of no less than 30 years (Subash and Sikka 2014, Rahmstorf et al. 2017, Asfaw et al. 2018). Understanding the gradual changes in temperature, rainfall,
wind speed and direction, atmospheric pressure and relative humidity are known as important indicators and essential parameters in the evaluation of the global hydro-meteorological response to climate change (Subash and Sikka 2014, Nguyen 2018). Different authors such as Aweda et al. (2022), Rahmstorf et al. (2017), Nguyen (2018), Zeng et al. (2016), Sun et al. (2015), Ilori and Ajayi (2015) have reported on the long-term change in the mean annual and seasonal temperature and precipitation which have become widespread in recent years. Some of these studies have focused on the global scale (Trenberth et al. 2003, Alexander et al. 2006, Liu 2011, Adler et al. 2017, Rahmstorf et al. 2017, Nguyen 2018, Ajayi and Ilori 2020), while others have examined the regional scale (Yao et al. 2010, Liu 2011, Williams et al. 2012, Bombardier et al. 2014, Ackerley et al. 2015, Priyan 2015, Sun et al. 2015, Zeng et al. 2016).

Globally, the average surface temperature between 1880 and 2012 which revealed a linear warming trend of 0.85°C is projected to rise between 1.4°C and 5.8°C by the year 2100 (Stocker et al. 2013, Keggenhoff et al. 2014, Driouech et al. 2020). Studies have shown that there is clear evidence of a rising surface temperature over the past decades (Stocker et al. 2013, Rahmstorf et al. 2017). Several studies also showed that rainfall pattern and trend have more variables depending on the region and specific atmospheric circulation phenomenon influenced by a warming climate that underlie the risk of floods, drought, loss of biodiversity and agricultural productivity (Trenberth 2011, Liu and Wu 2016, van Wilgen et al. 2016, Asfaw et al. 2018, Ayanlade et al. 2018, Aweda et al. 2021b). Most parts of southern Nigeria are coastal areas with high rainfall potential and a dense network of river tributaries (Ilori and Ajayi 2015). It was reported by Ilori and Ajayi (2015), that a little increase in rainfall amount duration or storminess could induce flood and threaten biodiversity in the region of southern Nigeria. However, a rise in temperature and delays in the onset of rain could create shifts in the cropping season and reduce agricultural productivity (Ilori and Ajayi 2015). Hence, to develop effective climate change mitigation and adaptation strategy, it is crucial to understand the annual and seasonal variation of the key indicators of temperature and rainfall (Ilori and Ajayi 2015,Aweda et al. 2021b).

Studies have shown a rising trend in air temperature across Nigeria (Ragatoa et al. 2018, Eresanya et al. 2018, Almazroui et al. 2020) with varying degrees of trends reported at annual and seasonal scales for different ecological regions (Abatan et al. 2016, Oluwatobi 2016, Abatan et al. 2018). There has been more variability in the general trend of precipitation, with rainfall observed to be increasingly less reliable during the crop-growing months in the rainforest and guinea savanna agro-climate zone of southwestern Nigeria (Ayanlade et al. 2018). Odjugo (2006) Observed a general decrease in rainfall across the country with a slight increase recorded in the coastal area using a fractional integration technique of Time Series Analysis. Gil-Alana (2012) Found a significantly positive time trend coefficient in rainfall across Nigeria by focusing on anomalies that masked seasonal variation. Oguntunde et al. (2011) Observed a statistically significant increase in rainfall and air temperature in Nigeria.
but with a decadal sequence of alternative increasing and decreasing trends in mean annual rainfall and air temperature. This article aims to analyze the significant effect of rainfall on air temperature, relative humidity, atmospheric pressure, wind speed and direction for forty-one years of measurement on a wide Nigeria strip in West Africa country. However, none of the previous studies has used MERRA-2 Re-Analysis data to study the effect of rainfall on other atmospheric parameters. Many authors have worked on rainfall trends in Nigeria, however, to the best of our knowledge, none has scientifically used Machine Learning to correlate the effect of rainfall on other atmospheric parameters. Therefore, this work adopted the use of Machine Learning to study the relationship between air temperature, pressure, relative humidity, wind speed and direction collected from the archive of HelioClim Satellite MERRA-2 data.

2 Material and Methods

2.1 Study location

The stations used for this research are different cities in Nigeria, representing various climatic zones of tropical Africa, which are partitioned into different coordinates of the north, south, east, and west (Figure 1, Table 1).
Table 1: The coordinate of the stations used.

<table>
<thead>
<tr>
<th>Stations</th>
<th>Longitude 'N'</th>
<th>Latitude 'E'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abuja</td>
<td>9.0723</td>
<td>7.4913</td>
</tr>
<tr>
<td>Ikeja</td>
<td>6.6059</td>
<td>3.3491</td>
</tr>
<tr>
<td>Ilorin</td>
<td>8.5000</td>
<td>4.5500</td>
</tr>
<tr>
<td>Maiduguri</td>
<td>11.8333</td>
<td>13.1500</td>
</tr>
<tr>
<td>Port Harcourt</td>
<td>4.7774</td>
<td>17.0134</td>
</tr>
<tr>
<td>Sokoto</td>
<td>13.0058</td>
<td>5.2475</td>
</tr>
</tbody>
</table>

2.2 Data collection and preparation

The data used in this study consist of monthly rainfall, air temperature, relative humidity, atmospheric pressure, wind speed and direction for five stations obtained from the archive of the HelioClim website of Soda (http://www.soda-pro.com) of MERRA-2 meteorological Re-Analysis data (Gelaro et al. 2017, Aweda et al. 2020a, Aweda et al. 2020b, Aweda et al. 2021a, Adebayo et al. 2022). The data of forty-one years spanning from 1980 to 2020 were obtained as a monthly average from January to December of every year in Comma Separated Value (CSV) data format.

2.3 Machine Learning Process

For this research, MATLAB and Python 3 were used for analysis of the data. The system analyzed the given meteorological data as input data by pre-processing and normalization for proper scaling. The processed data was further fed as input to multi-linear regression algorithms to predict the rainfall rate. The total dataset was divided into training and testing data. Training data comprises 70 percent of the total data and it was fed to the algorithms to derive a relationship between independent and dependent variables. Testing data, which comprises the remaining 30 percent was applied to the algorithms to compute values of the dependent variable or predicted values which were then compared with the actual values of rainfall, and the error rate was determined. The performance of models was evaluated on figures of merit to select a suitable model.

2.4 Mean Absolute Error (MAE)

One of the simple means of determining error magnitude is MAE. It comprises the mean absolute differences between the forecasted and the experimental values. This implies that higher values are not better than values closer to zero. In other words, values closer to zero are far better than higher values. They are:

\[
MAE = \frac{1}{Q} \sum_{q=1}^{Q} |PL_q - PL_q'| \quad (1)
\]

where \(PL_q\) is the prediction, \(PL_q'\) the true value and \(Q\) is the sample size.
2.5 Root Mean Square Error (RMSE)

The differences between the forecasted values of a model and the real values are of utmost importance in determining the RMSE. Therefore, the quadratic means of these differences are referred to as RMSE. The RMSE is most of the time if not all-time positive or zero, being this last value, the best one possible (but also an overfitting situation in many cases) (RMSE is also referred to as the square root of the mean of square errors). Consequently, the size of the error is important thereby penalizing large errors. It is, therefore becomes sensitive to outliers.

\[
RMSE = \sqrt{\frac{1}{Q} \sum_{q=1}^{Q} (PL_q - PL_q')^2}
\]

where \(PL_q\) is the prediction, \(PL_q'\) the true value and \(Q\) is the sample size.

2.6 R-Squared (R\(^2\))

The amount of the variance in the dependent variable that is predictable from the independent variable (R-Squared) explains the performance of a model when duplicating the experimental outcome (observed). Its value usually ranges between zero and one; other authors stated that R-Squared may take values from -\(\infty\) to 1. Generally, the closer the values are to 1, the better it is than values closer to zero or negative.

\[
R^2 = 1 - \frac{\sum_i(y_i - \hat{y}_i)^2}{\sum_i(y_i - \bar{y})^2}
\]

where \(R^2\) is the root mean square, \(y_i\) is the prediction and \(\hat{y}_i\) the true value.

3 Results and Discussion

3.1 Model performance

Performance metrics (error measures) are vital components of the evaluation frameworks in various fields. In Machine Learning Regression experiments, Performance Metrics are used to compare the trained model predictions with the actual (observed) data from the testing data set. Different models were used to predict the rainfall data (values) at different positions in the test dataset. These results were compared with observed data leading to the computation of the prediction error. Metrics employed to evaluate the performance of the predictors include MSE, RMSE, MAE, Speed of Prediction, R-Squared and Training Time.
3.2 Data Visualization

Before training algorithms with the dataset, there is the need to visualize the dataset to know the distribution pattern. A good distribution will enhance the performance of the trained model (Figure 2 a, b).
Table 2. Comparison of prediction accuracy of different predictors on the 20% test samples.

<table>
<thead>
<tr>
<th>Gaussian Process Regression Model</th>
<th>RMSE</th>
<th>MSE</th>
<th>MAE</th>
<th>$R^2$</th>
<th>Prediction Speed (obs/sec)</th>
<th>Training time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rational Quadratic GPR</td>
<td>0.51980</td>
<td>0.27019</td>
<td>0.38666</td>
<td>0.91</td>
<td>15000</td>
<td>7.9368</td>
</tr>
<tr>
<td>Squared Exponential GPR</td>
<td>0.52093</td>
<td>0.27137</td>
<td>0.38691</td>
<td>0.91</td>
<td>24000</td>
<td>1.8923</td>
</tr>
<tr>
<td>Matern 5/2 GPR</td>
<td>0.51799</td>
<td>0.26831</td>
<td>0.38525</td>
<td>0.91</td>
<td>17000</td>
<td>2.3701</td>
</tr>
<tr>
<td>Exponential GPR</td>
<td>0.52002</td>
<td>0.27042</td>
<td>0.39351</td>
<td>0.91</td>
<td>21000</td>
<td>3.2567</td>
</tr>
<tr>
<td>Optimized GPR</td>
<td>0.50342</td>
<td>0.25343</td>
<td>0.36976</td>
<td>0.92</td>
<td>26000</td>
<td>282.19</td>
</tr>
</tbody>
</table>

3.3 Skewed and normalization of data

Figure 2a shows the distribution of each of the features present in the dataset. Some data were skewed such as relative humidity and rainfall. Skewed data means that the tail region of the data may act as outliers for the model which can adversely affect the performance of the model especially the regression-based model. To attain a more evenly distributed dataset, we introduced a form of normalization technique called Min-Max Normalization. This is a linear transformation of the original dataset by scaling the data from 0 to 1 and it is given as:
\[ v' = \frac{v - \min_f}{\max_f - \min_f} (new\_\max_f - new\_\min_f) + new\_\min_f \]  \hspace{1cm} (4)

where \(v'\) is the new value of each entry in data, \(v\) is the old value of each entry in data, \(new\_\max_f\), \(new\_\min_f\) is the max and min value of the range (i.e boundary value of range required) respectively. Min-max normalization is one of the most popular ways to normalize data. For every feature, the minimum value of that feature gets transformed into a 0, the maximum value gets transformed into a 1 and every other value gets transformed into a value between 0 and 1.

The result of normalization is as shown in Figure 2b. Visualizing the relationship between the output variable (rainfall) and each input variable, as shown in Figure 3, reveals that while some input variables (relative humidity, rainfall, pressure, wind speed and direction) have a strong correlation with the output variable (rainfall), others have a noisy relationship which is not very clear. Thus, there is a need to use algorithms to understand and learn the hidden pattern for better prediction. The input variables and the output variable (rainfall) demonstrate that the output variable is known as a sub variable used for the determination of the result, which can serve as validation of the results.

The performance metrics of the different model were listed in Table 2. It is observed that optimized GPR performs better than other models but with lower prediction speed and training time.

![Predicted versus Actual Plot](image)

**Fig. 4.** Predicted versus Actual Plot (a) Rational Quadratic GPR; (b) Optimized GPR
The predicted versus the actual response plot was used to check model performance. This plot reveals the predictive ability of the regression model for different response values. A regression model is seamless if it has a true response equal to the predicted response. Consequently, all the data point must be on a diagonal line. The error of the prediction for a point is the vertical distance from the line to that point as reveal in Figures 4(a & b). It is a well-known fact that a model that gives small errors is a good one, therefore, the prediction is disperse closed to the line as seen in Figure 3b. The plot in Figure 4(b) shows fewer errors as compared to (a). Figures 5 (a & b) shows a relationship between MAE and RMSE for different models and it was observed that optimized GPR has a better performance as compared to others. Although the prediction time is longer as shown in Figure 4b.

5 Conclusions

This study investigated selected meteorological data prediction leveraging on Machine Learning Algorithm Approach over some stations in Nigeria. The algorithm of Machine Learning was explored using weather parameters such as rainfall, relative humidity, atmospheric pressure, wind speed and direction to predict rainfall rate over selected stations in Nigeria. The predicted response as against true response for the two models shows a linear graph passing through the origin which confirmed a perfect regression model where all the points lie on a diagonal line. Therefore, the relationship
between MAE and RMSE for the different models shows that optimized GPR has a better performance as compared to others. More so, visualizing the relationship between the output variable (rainfall) and each input variable, reveals that some input variables (relative humidity, rainfall, pressure, wind speed and direction) have a strong correlation with the output variables, while others have a noisy relationship, which is not very clear.

The results will aid in the prediction of weather by meteorological stations, which will include the prediction of climate change, which has been a significant change in Sub-Saharan Africa. However, when the results of this study are compared to those of other studies, it is clear that machine learning algorithms aid in proper prediction, as reported by Aweda et al. 2021a.

Acknowledgements

We hereby wish to express our profound gratitude to the HelioClim website MERRA-2 Meteorological Re-Analysis, for the provision of data for this research and Bowen University, Iwo, for the opportunity granted to conduct this research. We also appreciate the anonymous reviewers of RJS.

References

Segun Adebayo et al.

Meteorological prediction in Sub-Sahara Africa by Machine Learning


Segun Adebayo et al.  
Meteorological prediction in Sub-Sahara Africa by Machine Learning


