BOOTSTRAP AGGREGATED DECISION TREES FOR MODELING EVAPORATION PICHE USING OTHER METEOROLOGICAL FACTORS OVER ILORIN AND SOKOTO

*Are Stephen Olusegun[1], Alabi Nurudeen Olawale[1]

*Corresponding Author
[1]Department of Mathematics and Statistics, Federal Polytechnic Ilaro, Ogun State, Nigeria
*_stephen.are@federalpolyilaro.edu.ng, nurudeen.alabi@federalpolyilaro.edu.ng_

Abstract

The threat of climate change in recent times cannot be overemphasized particularly in developing economies largely due to the connection it has with national development issues. Nigeria like most emerging economies is highly susceptible to the impact of climate change because her economy is mainly dependent on income generated from the export of crude oil. One of the main meteorological or climatic factors associated with climate change is evaporation. This current work presents models inform of decision trees to study the relationships existing between evaporation and other meteorological factors such as relative humidity, solar radiation, sunshine hours, wind speed, temperature and rainfall over the ancient cities of Sokoto and Ilorin in Nigeria. These factors are generally understood to change with rising climatic changes in a place. Analysis of the fitted trees which was done using the recursive binary splitting (RBS), cost complexity pruning, and bootstrap aggregated (boosting) reveal that relative humidity is by wide margin and varying impact, the most important meteorological factor affecting evaporation piche in both cities. Its impact is slightly more prevalent in Ilorin than Sokoto.

Keywords: climate change, meteorological factors, recursive binary splitting, cost complexity pruning, boosting.

1.0 Introduction

Nigeria, like many other countries is exposed to climate induced dangers of desertification, erosion, flooding and other ecological problems which have impacts on the welfare of millions of people. A significant portion of the country's population rely majorly on agriculture a main occupation. Climate change has affected the agricultural sector in recent years. Also, climate change have affected human activities either directly or indirectly in many ways which include temperature changes, rainfall etc. Energy services are necessary inputs for every nation's development and growth just as it serves as the engine for sustainable economic growth and development. Generally speaking, the supply of energy entails the production/generation, transmission and distribution of electricity. Nigeria due to its strategic location is endowed with sizeable amount of energy resources such as thermal, hydro, solar, oil resources. Nigeria like most emerging economies is highly vulnerable to the impact of climate change because her economy is mainly dependent on income generated from production, processing and export of crude oil and related products. Building Nigeria's Response to

Climate Change (BNRCC, 2011) report asserted that hydro power generation is the energy source most probable to be influenced by climate change as it is sensitive to the amount of timing and geographical pattern of precipitation as well as temperature BNRCC (2011). The report also stated that a reduced flow in river and higher temperature reduces the capability of thermal electric generation as higher temperature also reduces transmission capacity. Also, excessive drought lead to higher evapo-transpiration that adversely affects water volume thereby reducing hydroelectricity capability. The effect of this drought on the power plants has led to a drastic reduction in the expected power supply from Kainji Dam. Evaporation is one of the major processes in the hydrological cycle. Evaporation is perhaps the most difficult component because of complex interactions of the component of the land-plant- atmosphere system. Evaporation depends on the supply of heat energy and vapour pressure gradient, which in turn depend on meteorological factors such as temperature, wind speed, atmospheric pressure, solar radiation, quality of water and the nature and shape of evaporation surface. These factors also depend on other factors, such as geographical location, season, time of day and so on.

1.1*Areas under study*

The ancient city of Ilorin, the capital of Kwara state, is located between latitude 8^0 24’ and 8^0 36’ north of the equator and between longitude 4^0 10’ and 4^0 36’ east of the Greenwich meridian. Ilorin is a transition zone between the deciduous wood land of the south and dry savanna of north. The city has an approximate area of 150 sq.km and population of about 777,667 (2006 census). Ilorin has a humid tropical climate which is characterized by wet and dry seasons. The temperature in the city is uniformly high throughout the year and open air insulation can be very uncomfortable during the dry season. The mean monthly temperatures are usually very high varying between 25.1^0C in August and 30.3^0C in March. The area’s daytime range of temperature is also high. Rainfall in Ilorin is produced by tropical continental air mass and it exhibits great variability both temporarily and spatially with mean annual rainfall of about 1200mm; relative humidity in the city during the wet season is between 75 per cent to 80 per cent while in the dry season it is about 65 per cent.

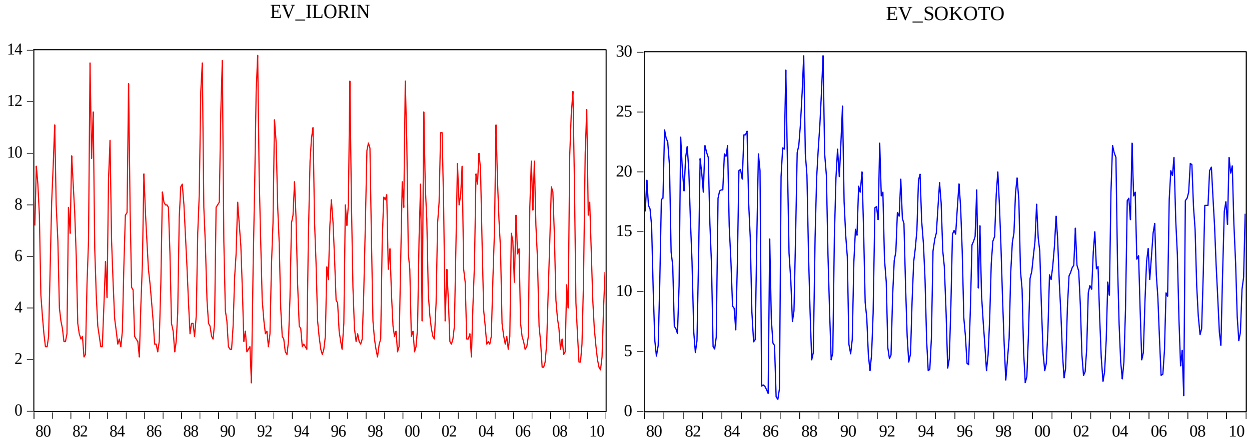


FIGURE 1. *Time plots on evaporation piche over Ilorin and Sokoto.* ***Left-hand Panel****: Time plot on evaporation piche over Ilorin.* ***Right-hand Panel****: Time plot on evaporation piche over the city of Sokoto. Both plots show rather random pattern with regular disturbances in the distribution of the meteorological factor. These time plots also reveal generally higher evaporation piche over Sokoto than Ilorin.*

Sokoto is another of the ancient cities located in the North West region of Nigeria with global coordinates 13^{o}04'N 5^{0}14'E and 427,760 (2006 census). It is also situated in the Sahel Savannah with an average temperature of 28.3^{0} C (82.9^{0} F). Its temperature ranges between 40^{0}C and 45^{0}C during the daytime. It covers an are of land of about 145 sq. km. The city lies in the lullemmeden Basin also refer to as the Sokoto Basin. This area consist of rising and falling plain with an average elevation ranging between 250mm and 400mm above sea level which is occasionally disturbed by low mesas (Obaje, 2009). The climate is broadly classified into wet and dry seasons. The wet season starts early in year with a mean annual rainfall ranging between 500 mm and 1,300 mm. The dry season is predominantly hamattan with dry, cold, and fairly dusty wind from the Sahara desert. The main economic activity of the city is agriculture (both crop and animal husbandry). Cash crops such as wheat, cotton and vegetables are common. Other crops cultivated include millet, guinea corn, maize, rice, potatoes, cassava, groundnuts and beans.

1.2 Some related works

Wang, (2006) research on the influence of climate change on evaporation started in late 1970s. It was generally considered that the temperature is an important influencing factor for evaporation. Xie (2009) used partial correlation coefficients to describe the linear relationship between two variables and concluded that this is more reasonable and reliable for practical application. Partial correlation coefficients can relatively reflect the degree of dependency between evaporation and meteorological factors by integrating out all other affecting variables. Shen et al. (2009) used the Mann-Kendall test to detect the variation trends of E601 pan evaporation over China from 1960 to 2006. The results showed that during the last 50 years, temperature has presented a significantly increasing trend while E601 pan evaporation shows a decreasing trend for most climate zones of China before the late 1990$_{s}$. The result reveals that daily temperature range, sunshine duration and average wind speed correlated well within the E601 pan evaporation, and significantly decreasing trends of these influencing factors are main reason explaining the decreasing rates of E601 pan evaporation. Kay and Davies (2008) investigated the differences in the estimation of monthly potential evaporation in Britain when using outputs from either General Circulation Models (GCM$_{s}$) or (RCMs). They utilized the Penman-Monteith model and a simple temperature based model in the estimation of evaporation for the current climate. The results were compared with a dataset derived from a modified Penman-Monteith formulation using meteorological factors. According to their results, RCM'$_{s}$ outputs are able to generate monthly evaporation rates that show much closer agreement with evaporation rates derived from observed data as compared to GCM'$_{s}$ outputs.

Roderick et al. (2009) performed a study on global pan evaporation trends and found that there is an overall declining trend over a 30 to 50 year period. More specifically, they found most analyzed sites to range from -1 to -4mm per year after a 30-year period. Roderick (2009) showed that there has been an overall decrease of 4.8W$_{m}^{-2}$ over the past 30 years. Shih (1984) employs nine meteorological variables to estimate pan evaporation using multiple regression with the ordinary least square analysis or the ridge regression analysis. The Thornthwaite method for estimation of evaporation and ET is basically dependent on the average monthly temperature, number of days in the month, and number of hours between sunrise and sunset. Kisi (2009) compared the performances of three different ANN techniques and it was found that the Multilayer Perceptron (MLP) and Radial Basis Neural Network (RBNN) computing techniques could be employed successfully to model the evaporation process using the available climatic data.

Xu et al. (2006) conducted an analysis of four meteorological variables in the Yangtze basin of china by reporting a decreasing trend of evapo-transpiration in the region. Evaporation for the corresponding period was also reported having a decreasing trend. Out of the four meteorological parameters analyzed, air temperature and relative humidity showed an increasing trend while wind speed and net radiation showed a declining trend. The decrease in solar radiation was suggested to be most likely from a decrease in global radiation and may also be due to pollution. Also, the work cited reports of decreasing mean cloud cover over China. The increase in air temperature was reported to be consistent with global warming reports. The results showed that the ensemble network Neural Network Autoregressive Model with Exogenous Inputs (NNARX) is better than the Artificial Neural Network (ANN) and Marciano method. The models with inputs such as of wind speed and vapor pressure performed much better than the ones with temperature and dew point. Peterson et al. (1995) proposed a model on pan evaporation data (1945-1990) from the eastern and western United States, Europe, Middle Asian and Siberian regions of the former Soviet Union showed a significant decline of meteorological factor. The largest change reported was 97mm increase in a warm season (May-September) for the western United States during the past 45 years and the study suggested that a feature of recent climate change includes a decrease in potential evaporation (ET_p). The decrease in pan evaporation was attributed to a decrease in the diurnal temperature range and an increase in low cloud cover. Chang et al. (2010) proposed a Self-Organizing Map Neural Network (SOMN) to assess the variability of daily evaporation on meteorological variables. Their work demonstrated that the topological variables and the networks could well estimate the daily evaporation.

Kim et al. (2012) applied Multi-layer Perceptron Neural Networks (MLP), Generalized Regression Neural Networks (GRNN) and Support Vector Machine Neural Networks (SVMNN) to estimate E_p in temperature and climatic zones and the results indicated that these ensemble models performed better than the empirical Linacre model and ANN. Jun and Hideyuk (2004) reported that pan evaporation is in a declining trend throughout Japan for all seasons, based on 34 years of evaporation data from 13 sites. The cause for the few exceptional stations not following the trend was attributed to local urbanization influence. Stated factors for declining pan evaporation were increasing vapor deficit and increasing terrestrial evapo-transpiration. Wang et al. (2017) in their paper compared the performances of six heuristic computation techniques based on Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coefficient of determination (R^2) and found that amongst MLP, Least Squares Support Vector Machine (LSSVM), Fuzzy Genetic (FG), Adaptive Neuro-Fuzzy Inference System (ANFIS) techniques, Multiple Linear Regression (MLR), as well as the Stephens-Stewart (SS) methods, the MLP performed best while MLR performed the least. Herch and Burn (2005) applied the Meyer’s formula to analyze trends of evaporation over Canada for 30, 40 and 50 years period. The work produced varying trends of gross evaporation and pan evaporation especially June, July, October. Furthermore, a significant decreasing trend in annual evaporation was reported. Kisi et al (2005) investigated the accuracy of LSSVM, Multivariate Adaptive Regression Splines (MARS) and M5 model tree (M5 tree) in modeling evaporation using local input and output data while the MARS model performed better than the LSSVM model in the case of without local input and outputs. Hess (1998) analysis of evaporation measurements (1964-1998) in the central coastal plains of Israel showed a small but statistically significant increase in pan evaporation from screened class pans.

2.0 Methods, Empirical Analyses and results

The meteorological data over Ilorin and Sokoto are samples with the same size n = 372 observations on evaporation piche (**ev**), solar radiation (**sr**), wind speed (**wd**), sunshine hours (**sh**), rainfall (**rf**), relative humidity (**rel**) and change in temperature (**te**). These data were extracted from the database of Nigeria Meteorological Agency (NIMET). We divided the monthly data into training and test datasets. The former was used to train the decision trees while the latter to test the performance of the fitted tree models. Just like in a tree analogy, the method of decision trees for regression and classification comprises leaves generally referred to as the *terminal nodes* and the points along which the regressor space is split called the *internal nodes*. The internal and terminal nodes are connected by sections called the *branches*. These methods are appropriate if the relationships between the response and the regressors are complex and highly non-linear. We draw our inspiration from the procedure outlined by Hastie et al. (1996) which involves dividing the regressor space $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6]$ where $\boldsymbol{x}_1$ = **sr**, $\boldsymbol{x}_2$ = **sh**, $\boldsymbol{x}_3$ = **wd**, $\boldsymbol{x}_4$ = **rf**, $\boldsymbol{x}_5$ = **te**, $\boldsymbol{x}_6$ = **rel** into m distinct and disjoint box-shaped regions $Z_1, Z_2, ...Z_m$ in which the response averages in each box are as different as possible. The splitting procedures describing the regions are related to each through a binary tree done recursively. The prediction is then done by determining the region Z_j in which an observation falls into and using the mean of the response values of the training observations in that region Z_j as the predicted value for that observation. The response v = **ev** was modeled as a constant c_j in each region as

$$f(X) = \sum_{j=1}^{m} c_j . 1_{(x \in Z_j)} \qquad 1.0$$

The regions are determined such that the residual sum of squares, RSS, given in equation 1.1 is minimized

$$RSS = \sum_{j=1}^{m} \sum_{i \in Zj} (v_i - \hat{v}_{Zj})^2 \qquad 1.1$$

where $\hat{v}_{Zj}$ is the mean response (evaporation piche, **ev**) for training observations in the j^{th} region. We employed an efficient procedure called the *recursive binary splitting* (RBS) in growing our decision tree rather than generate a rather costly and numerically infeasible j partitions of the regression space.

2.1 Recursive Binary Splitting (RBS) on the fitted Evaporation Piche tree model

We adopted the *recursive binary splitting* (RBS, hereon) which start with a single region at the top of the tree and gradually perform splitting in an optimal fashion at each step of the tree construction. The RBS algorithm select with an explicit halting condition, a regressor $\boldsymbol{x}_j$ in $\boldsymbol{X}$ and a cutoff u such that the regressor space is split into two regions

$$Z_1 = \{X \mid x_j < u\} \text{ and } Z_2 = \{X \mid x_j \geq u\}$$

with the ultimate aim of reducing the RSS in equation 1.1. This implies that we attempt to generate the value of *j* and *u*, which minimize equation 1.2

$$\sum_{i:x_i \in z_1(j,u)} (v_i - \hat{v}_{z_1})^2 + \sum_{i:x_i \in z_2(j,u)} (v_i - \hat{v}_{z_2})^2 \qquad 1.2$$

where $\hat{v}_{Z_1}$ is the mean response (**ev**) for training observations in the $Z_1(j, u)$ and $\hat{v}_{Z_2}$ is the mean response for training observations in the $Z_2(j, u)$ regions. This process was repeated in the subsequent steps minimizing RSS in each step. This resulted in the tree in **Figure** 2 with the value of *j* and *u* that minimizes the RSS in equation 1.1 are 6 and 77.5 per cent respectively for Ilorin. Similarly, j = 6 and u = 53.5 per cent for Sokoto. That is relative humidity (**rel**) is the regressor at the top of the tree used for the initial split such that

$$Z_1 = \{X \mid rel < 77.5\%\} \text{ , } Z_2 = \{X \mid rel \geq 77.5\%\}$$

and

$$Z_1 = \{X \mid rel < 53.5\%\} \text{ , } Z_2 = \{X \mid rel \geq 53.5\%\}$$

for Ilorin and Sokoto respectively. For the evaporation piche model for Ilorin the value of RSS = 491, Root Mean Square Error (RMSE) = 1.657 and the number of terminal nodes = 7. Similarly, the RSS and RMSE in the evaporation piche model for Sokoto are 4,124.16 and 4.8 respectively. The splitting were terminated as soon as we have not more than 5 observations in each region **Table** 1. The unpruned trees used up about 88 per cent and 95 per cent of the training observations respectively.

Table 1. *Summary of RBS on the two Evaporation Piche Tree Models*

ILORIN					SOKOTO				
Node	**Split**	**Number of Observations**	**RSS**	**Predicted Evaporation**	**Node**	**Split**	**Number of Observations**	**RSS**	**Predicted Evaporation**
1	Root	186	1502	5.209	1	root	186	7194	12.49
2	rel < 77.5	94	482.9	7.462	2	rel < 53.5	113	2559	16.21
3	rel > 77.5	92	54.62	2.907	3	rel > 53.5	73	645.7	6.729
4	rel < 48	9	16.05	11.710*	4	rel > 17.5	27	293.4	19.3
5	rel > 48	85	287.1	7.012	5	rel > 17.5	86	1927	15.24
6	rel < 80.5	13	4.169	3.992*	6	rel < 68	25	287.8	9.184
7	rel > 80.5	79	32.6	2.728*	7	rel > 68	48	128.7	5.45*
10	rf < 31.45	50	161.9	7.81	8	wd < 7.1	7	63.91	16.11*
11	rel > 31.45	35	47.81	5.871*	9	wd > 7.1	20	133.3	20.42*
20	te < 15.05	45	124.8	8.007	10	sh < 8.95	52	1072	14.11
21	te > 15.05	5	19.71	6.040*	11	sh > 8.95	34	685.8	16.98
40	rel <57.5	5	10.64	9.900*	12	wd < 6.45	8	49.27	6.625*
41	rel > 57.5	40	94	7.770*	13	wd > 6.45	17	161.5	10.35*
NA	NA	NA	NA	NA	20	sh < 7	9	89.19	17.19*
NA	NA	NA	NA	NA	21	sh > 7	43	879.4	13.47*
NA	NA	NA	NA	NA	22	wd < 8.85	23	271	15.36*
NA	NA	NA	NA	NA	23	wd > 8.85	11	229.2	20.35*

Terminal node (leave)* **Source: *Authors' Computation using* ***R*** *language*

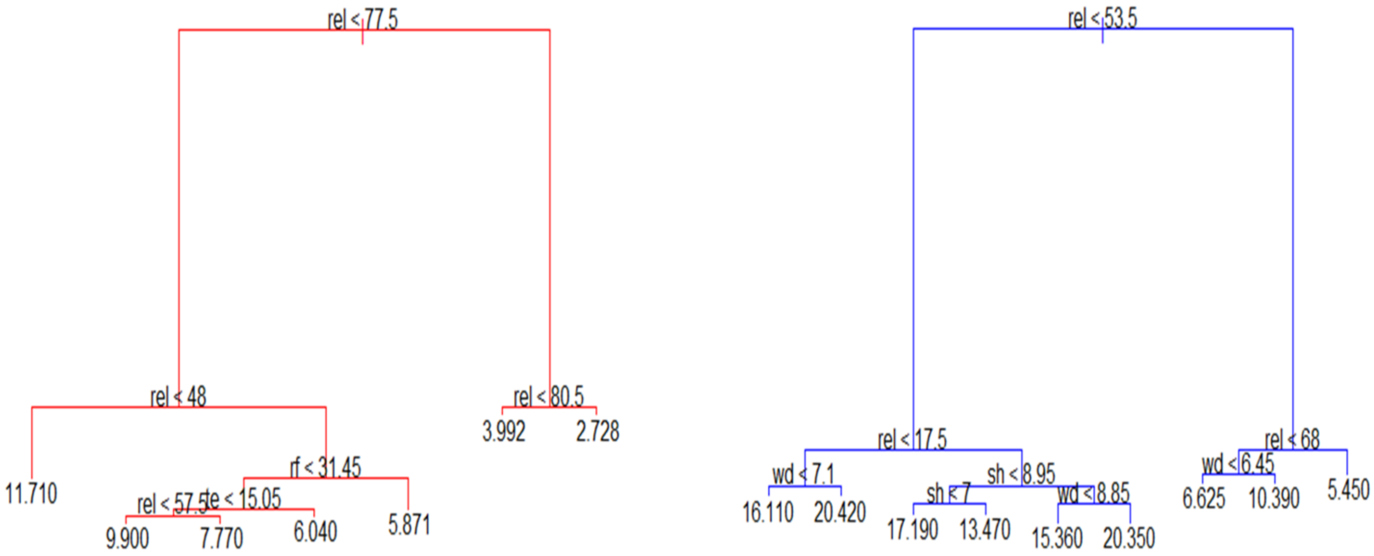


FIGURE 2: *These tree models show that relative humidity* ***(rel)*** *is the regressor in the evaporation piche tree model to minimize the RSS in equation 1.1. Splitting was done using the recusrsive binary algorithm. At each step, a regressor is selected for binary splitting. At a given internal node the left-hand branch is represented by* $\boldsymbol{x}_j < q_k$ *resulting from the split and* $\boldsymbol{x}_j > q_k$ *indicates the right hand branch.* ***Left hand panel:*** *Unpruned decision tree with 7 leaves and 6 internal nodes. On the Ilorin data, at the top of the tree, the split resulted in two branches in which the left hand branch corresponds to* ***rel*** *< 77.5 per cent and the right hand branch corresponds to* ***rel*** *≥ 77.5 per cent. The the entire twofold tree splitting process resulted in relative humidity* ***rel****, rainfall* ***rf*** *and temperature* ***te*** *as internal nodes and evaporation piche as the terminal node.* ***Right hand panel****: Unpruned decision tree with 9 leaves and 8 internal nodes. This represent the decision tree model for Sokoto with split at the top of the tree corresponding to* ***rel*** *< 53.5 per cent on the left hand branch and otherwise on the right hand branch. This tree has relative humidity, wind speed and sunshine hours as the internal nodes and evaporation piche as the terminal node. Comparing the values of predicted evaporation piche for the two cities revealed that the Sokoto is expected to have higher evaporation piche than Ilorin. Splitting ensures simplicity and ease of interpretability of the regression tree model.* ***Source****: Authors' Computation using* ***R*** *language.*

This procedure of RBS resulting in **Figure** 2 produced good predictions on the two training datasets but eventually overfitting the evaporation piche data. One problem is that this might lead to a very poor performance of the decision tree models on the two test datasets. We were able to achieve a better test performance by generating smaller trees that contain fewer splits using *cost complexity pruning*. One benefit of this pruning method is that we were able to fit evaporation piche tree models with lower variance but slightly higher bias in a manner that utilizes only few number of subtrees unlike the cross-validation and the validation set pruning for the two cities. The cost complexity pruning involves a positive tuning parameter α such that for every value of this quantity, there exists a subtree $T \subset T_{o,}$ for which equation 1.3 is as small as possible.

$$C_\alpha |T| = \sum_{m=1}^{|T|} n_m Z_m(T) + \alpha |T|$$

$$where,$$

$$Z_m(T) = \frac{1}{n_m} \sum_{x_i = z_m} (v_i - \hat{v}_m)^2 \qquad 1.3$$

$$\hat{v}_m = \frac{1}{n_m} \sum_{x_i = z_m} v_i$$

$$n_m = \# \{ x_i \in Z_m \}$$

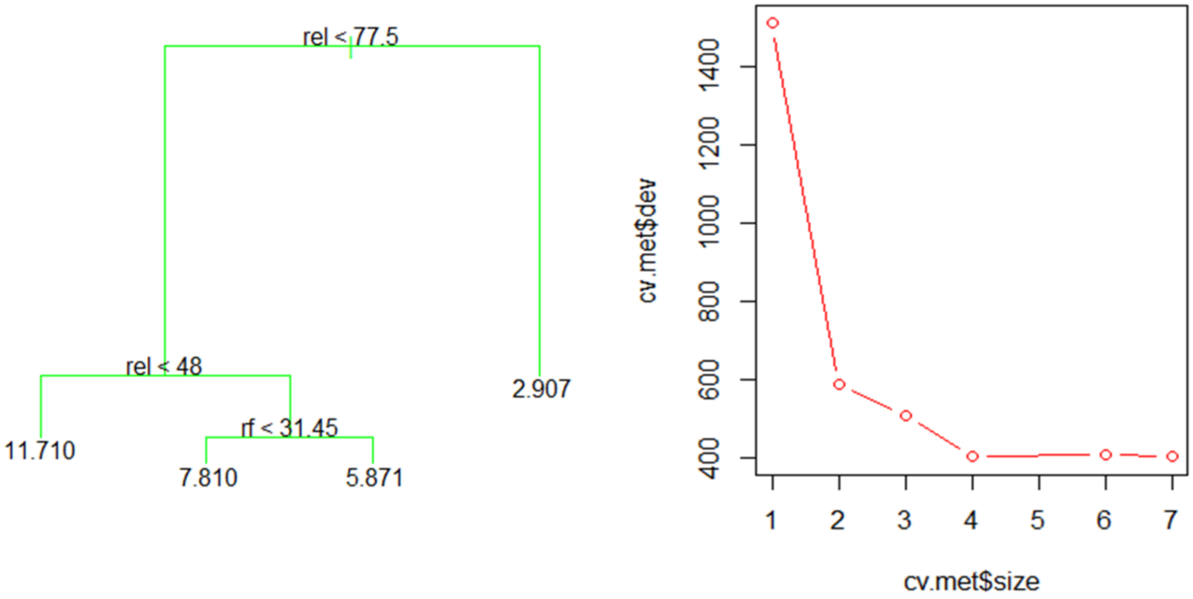
The quantity |*T*| is the number of leaves on the decision tree *T*, Z_m is the region relating to the *m*th leave and $\hat{v}_m$ is the mean of the training dataset's evaporation piche (**ev**) corresponding to the region Z_m. Increasing the value of α from zero in equation 1.3 prunes the branches and controls the tradeoff between the complexity of the regression subtree and its goodness of fit on the training datasets. We employed the 10-fold cross validation to determine the values of the positive tuning parameter and a most complex tree. We computed a tuning parameter α = 18.785 with an RSS value of 403.652 using cross validation for the data on the city of Ilorin. Similarly, for the city of Sokoto, α = 3,989.50 with an RSS value of 3,714.67. **Table** 2 shows the result of the cost complexity pruning for various values of α at each split on the two decision tree models. We set the value of the best number of terminal nodes to 4 and 6 for Ilorin and Sokoto respectively, which were calculated using the 10-fold cross-validation approach. These pruned trees were generated from large trees on the training datasets containing 186 observations and varying the nonnegative tuning parameter α in equation 1.3.

Table 2: *Cost Complexity Pruning on Regression Tree Model*

ILORIN					SOKOTO				
Node	Split	Number of Observations	RSS	Predicted Evaporation	Node	Split	Number of Observations	RSS	Predicted Evaporation
1	root	186	1502	5.209	1	root	186	7194	12.49
2	rel < 77.5	94	482.9	7.462	2	rel < 53.5	113	2559	16.21
3	rel > 77.5	92	54.62	2.907*	3	rel > 53.5	73	645.7	6.729
4	rel < 48	9	16.05	11.710*	4	rel > 17.5	27	293.4	19.3*
5	rel > 48	85	287.1	7.012	5	rel > 17.5	86	1927	15.24
10	rf < 31.45	50	161.9	7.810*	6	rel < 68	25	287.8	9.184*
11	rf > 31.45	35	47.81	5.871*	7	rel > 68	48	128.7	5.45*
NA	NA	NA	NA	NA	10	sh < 8.95	52	1072	14.11*
NA	NA	NA	NA	NA	11	sh > 8.95	34	685.8	16.98
NA	NA	NA	NA	NA	22	wd < 8.85	23	271	15.36*
NA	NA	NA	NA	NA	23	wd > 8.85	11	229.2	20.35*

Terminal node (leave)* **Source: *Authors' Computation using **R** language*

Cost Complexity Pruning of Decision Trees on Evaporation Piche using other meteorological factors



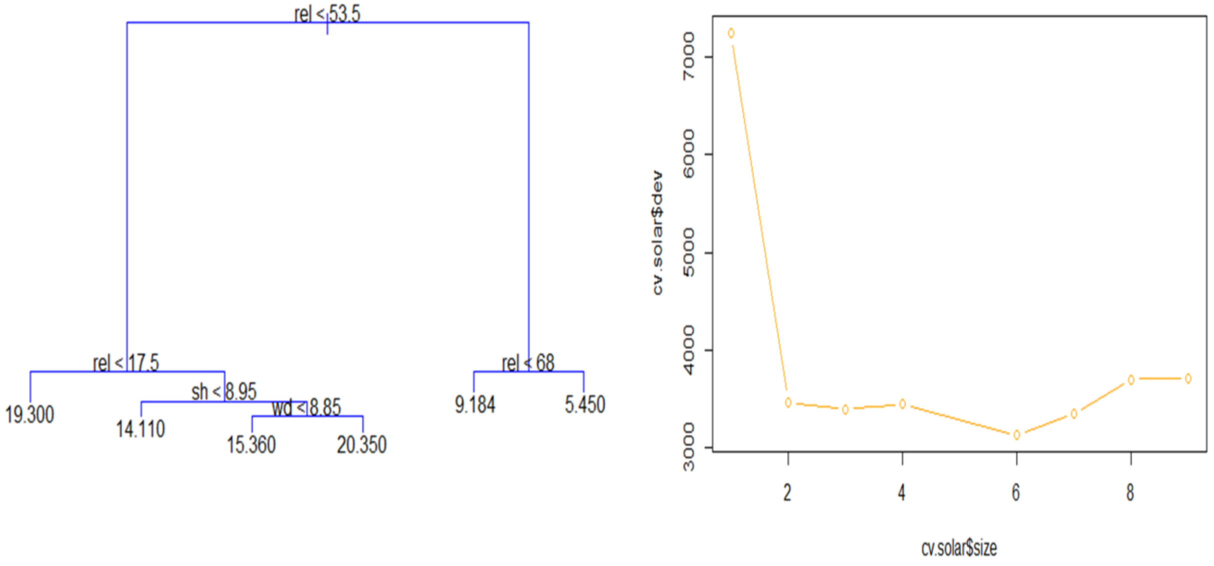


FIGURE 3. *Analysis of cost complexity pruning on the fitted decision trees of* ***Figure*** *2.* ***Upper Left panel****: Prunned tree with 4 terminal nodes, 3 internal nodes for Ilorin. At the top of the tree, the split resulted in two branches in which the left-hand branch corresponds to* ***rel*** *< 77.5 per cent and the right-hand branch corresponds to* ***rel*** *≥ 77.5 per cent.* ***Upper Right panel****: The result of 10-fold cross validation showing the cross-validation error (cv.met$dev) as a function of the terminal nodes (cv.met$size) for Ilorin.* ***Lower Left Panel****: This tree has 6 terminal nodes and 5 internal nodes after pruning using 10-fold cross validation with left hand branch corresponding to* ***rel*** *< 53.5 per cent and the right hand panel* ***rel*** *≥ 53.5 per cent for Sokoto.* ***Lower Right Panel****: The result of 10-fold cross validation showing the cross-validation error (cv.met$dev) as a function of the terminal nodes (cv.met$size) for the city of Sokoto.* ***Source****: Authors' Computation using* ***R*** *language.*

The pruned trees in **Figure** 3 reveal that out of the six regressors, relative humidity **rel** is the first meteorological factor for predicting evaporation piche over both cities. The strength of these decision trees lie in their interpretability. Specifically, for the model on Ilorin data, using 94 training observations, given that **rel** is greater or equal to 77.5 per cent, the mean of evaporation piche is 2.907 ml which represent the first terminal node. The fitted regression tree was further split by **rel** given that the relative humidity at the top of the tree is less than 77.5 per cent. At this step, if **rel** is less than 48 per cent, a total of 9 training observations produced a mean evaporation piche of 11.710 *ml*. Otherwise another split was carried out which resulted in two terminal nodes. These two leaves are 7.81 *ml* if rainfall (**rf**) is less than 31.45 *mm* and 5.871 *ml* otherwise. A total of 50 and 35 training observations were used up in calculating these two mean responses of the evaporation piche.

Using the result of this pruned tree, we predicted the evaporation piche for the 186 test observations to determine its performance. The RMSE of the pruned model using test dataset is 1.482 *ml*. Similar interpretations apply to the model on Sokoto data with an RMSE of 4.576 *ml* on the test dataset.

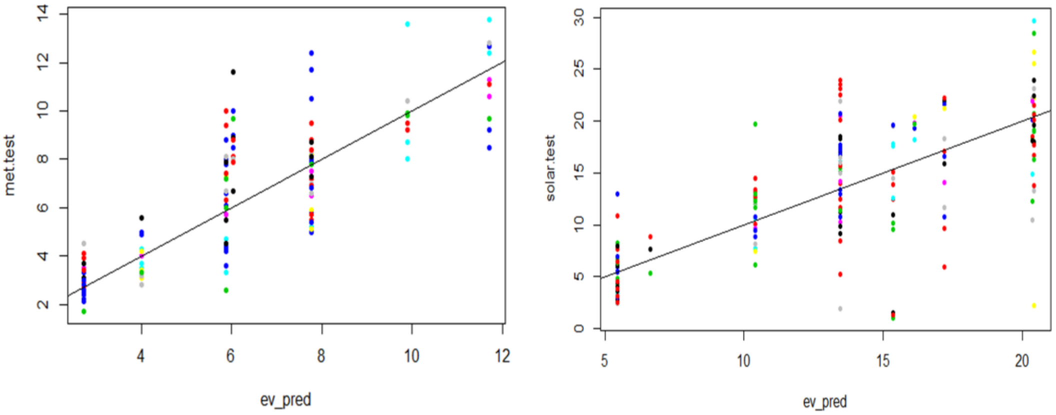


FIGURE 4. *The scatter plots on the residuals of the optimally-pruned regression trees using the test datasets. Left-hand panel: residual plot on evaporation piche model for the city of Ilorin. Right-hand panel: residual plot on the evaporation piche model for Sokoto city. These plots show the higher residual values for Sokoto. The straight line through the dotted points is the trend line.* ***Source****: Personal Computation using* ***R*** *language.*

2.2 Boosting the fitted Evaporation Piche decision tree models

Decision trees are known to present some setbacks in terms of predictive accuracy (higher variance) causing instability and lack of robustness to changes in data. Furthermore, the pruned trees using the cost complexity pruning performed poorly on the test datasets as revealed by the RMSE values. Therefore, we employed boosting to improve the accuracy of the predictions generated by the regression tree models. Boosting is a bootstrap aggregation procedure used in reducing the variance of a statistical learning method such as our fitted regression trees. Averaging a set of observations reduces the variance at the expense of higher bias. Here the bootstrapping involves taking single training dataset, we successively grow G = 40,000 different bootstrapped regression trees sequentially by utilizing information from the residuals ($\boldsymbol{r}_i$) of the previously grown trees. That is these trees were fitted using the residuals of previously grown trees rather than the original response values $\boldsymbol{v}_\text{i}$. Using G = 40,000 subtrees, we computed

$$\hat{f}^{1}(x),\ \hat{f}^{2}(x)................\hat{f}^{40,000}(x)$$

based on the following algorithm.

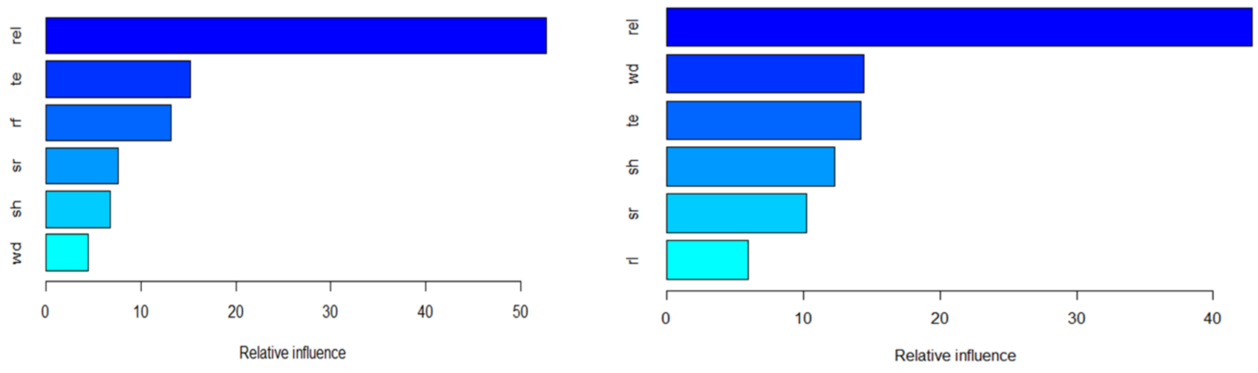
2.1.1 Algorithm on boosting for Evaporation Piche decision tree models

```
1. Set $\hat{f}(x) = 0$ and $r_i = v_i$ for all $i$ in the training dataset
2. Fit $g$ = 1, 2, 3,…………,40,000, repeat:
    a. Fit a tree $\hat{f}^{g}$ with $\boldsymbol{b}$ splits to the training data ($x$, $r$)
    b. Update $\hat{f}$ by adding in a shrunken version of the new tree:
           $\hat{f}(x) \leftarrow \hat{f}(x) + \delta\hat{f}^{g}(x)$
    c. Update the residuals,
           $r_i \leftarrow r_i \leftarrow \delta\hat{f}^{g}(x_i)$
3. Output the boosted model,
           $\hat{f}(x) = \sum_{g=1}^{40,000} \delta\hat{f}^{g}(x)$
```

Because these trees are grown deep and unpruned, they possess higher variance but low bias. The boosting algorithm terminated with G = 4,074 and 3,234 trees aggregated for Ilorin and Sokoto models respectively. In order to improve the performances of our boosted models by minimizing the variance, we set the shrinkage parameter $\boldsymbol{\delta}$ = 0.0159 and 0.001 for Ilorin and Sokoto respectively. The shrinkage parameter $\boldsymbol{\delta}$ (otherwise called the learning rate of the boosting process), alongside the interaction depth parameter ***b*** (the number of splits on each of the ***g*** tree) and the number of grown trees ***g***, constitute the tuning parameters of the boosting procedure. We used the depth parameter value of ***b*** = 1 which implies that at each split, only one variable was used. RMSE of the boosting procedure using 15-fold cross-validation on test data set is 1.181 *ml* which is about 64 per cent improvement over the cost complexity pruned model for Ilorin. While the RMSE = 4.307 *ml* accounting for more than 11 per cent improvement over the cost complexity pruning. We were able to generate the relative influence of the regressors in the evaporation piche tree model using the boosting procedure **Table** 3 and **Figure** 5.

Table 3. Analysis of regressor relative influence on evaporation piche over Ilorin and Sokoto

Regressor	**Relative Influence (%)**	
	Ilorin	**Sokoto**
Relative Humidity (**rel**)	52.7	42.82
Temperature (**te**)	15.22	14.48
Rainfall (**rf**)	13.14	5.93
Solar radiation (**sr**)	7.66	10.3
Sunshine hours (**sh**)	6.81	12.13
Windspeed (**wd**)	4.47	14.32



***Source**: Authors' Computation using **R** language*

FIGURE 5. *Relative Influence of each meteorological factor is displayed in these plots.* ***Left-hand Panel**: The plot reveals that relative humidity is largely the most important factor (accounting for more than 50 per cent of the total variability) affecting evaporation piche in Ilorin. This is followed by temperature (about 15 per cent), then rainfall (about 13 per cent). Wind speed has the least impact (less than 5 per cent of the total variability) on evaporation piche in the city. Solar radiation, sunshine hours and wind speed accounted for less than 10 per cent of variability in evaporation piche over Ilorin.* ***Right-hand Panel**: This panel shows the plot on the relative importance of these factors on evaporation piche over Sokoto. Whilst Relative humidity accounted for more than 40 per cent, wind speed and temperature accounted for about 15 per cent of the total variability in evaporation piche over Sokoto. Only rainfall accounted for less than 10 per cent of total variability (Table 3).* ***Source**: Authors' Computation using **R** language.*

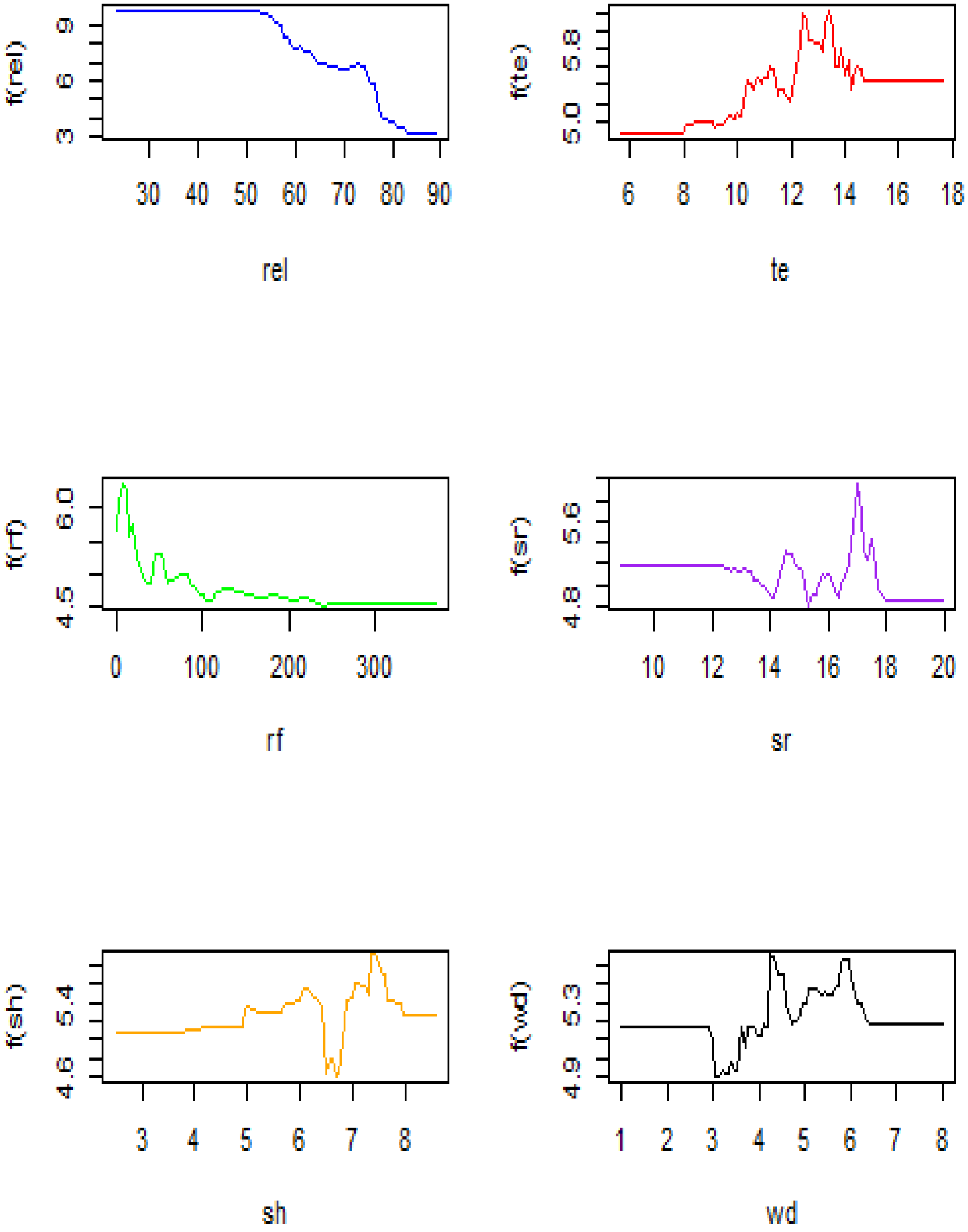

f(rel)
3
6
9
30
40
50
60
70
80
90
rel
f(te)
5.0
5.8
6
8
10
12
14
16
18
te
f(rf)
4.5
6.0
0
100
200
300
rf
f(sr)
4.8
5.6
10
12
14
16
18
20
sr
f(sh)
4.6
5.4
3
4
5
6
7
8
sh
f(wd)
4.9
5.3
1
2
3
4
5
6
7
8
wd

FIGURE 6. *Partial Dependence plots showing the marginal effects of the meteorological factors on evaporation piche over Ilorin. These plots are the marginal effects of the corresponding meteorological factor integrating out the other factors in the fitted decision tree model. These plots show that evaporation piche declines steadily with increase in relative humidity. Similarly, as rainfall rises in the city, evaporation piche declines. However, rising temperature resulted in increased evaporation piche.* ***Source****: Authors' Computation using* ***R*** *language.*

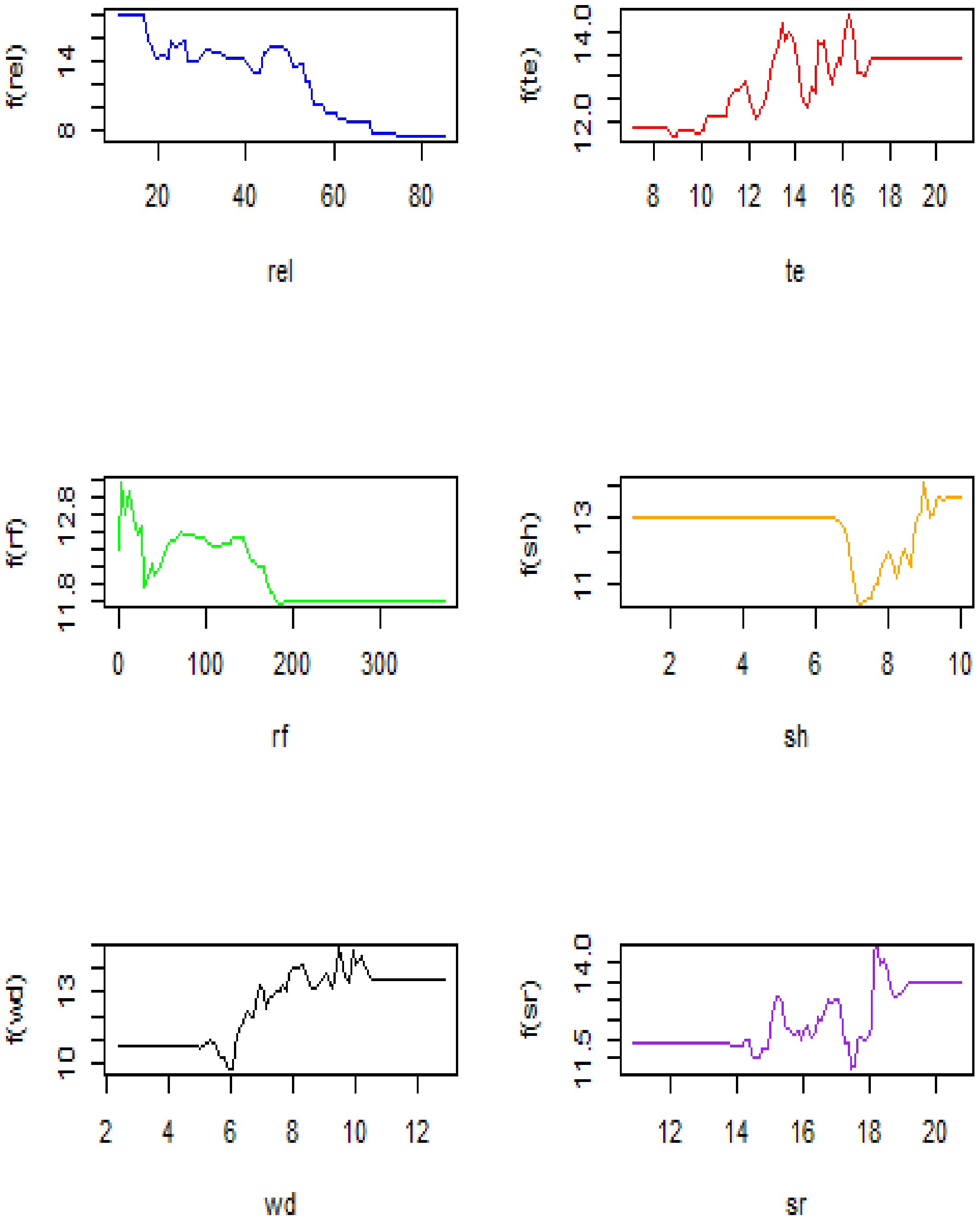


FIGURE 7. *Partial Dependence plots showing the marginal effects of the meteorological factors on evaporation piche over Sokoto. The marginal effects of the corresponding meteorological factor was done* *integrating out the other factors in the fitted decision tree model the same manner as Ilorin tree model. These plots show that evaporation piche declines steadily with increase in relative humidity and rainfall. On the contrary, as temperature, wind speed, sunshine hours and solar radiation rises over Sokoto, evaporation piche rises too.* ***Source****: Authors' Computation using* ***R*** *language.*

3.0 Conclusion

Decision trees for regression have been used to study the relationships between evaporation piche and six other meteorological factors influencing it over Ilorin and Sokoto. The analysis of the trees was done using the Recursive Binary Splitting (RBS), cost complexity pruning, and boosting. Recursive Binary Splitting produced trees with seven and nine leaves or terminal nodes for Ilorin and Sokoto respectively. Though, this single tree produced low RMSE for the models on the training datasets, their performances on the test datasets were weak. Relative humidity minimized the residual sum of squares (RSS) in the two decision tree models. Therefore, this meteorological factor was used as the initial splitting variable at the top of the trees. We applied the cost complexity pruning (CCP) to trim down the number of terminal nodes (leaves) and improve the predictive capacity alongside the interpretability of our models. These models resulted in lower model RMSEs for the test datasets and fewer terminal nodes. The Cost Complexity Pruning procedure was based on varied values of a tuning parameter which was used to control the tradeoff between the complexity of the models and it overfitting the models. We employed boosting to further improve the predictive capacity of the postulated decision tree models. Boosting is a procedure that involve growing large number of separate trees sequentially using the residuals from previously grown tree as the response in the new tree. The results of boosting revealed improved performances of the regression tree models using the test datasets in terms of lower RMSE or accuracy. Partial dependence analyses of the regressors in terms of marginal effects indicate that evaporation piche rises with rising temperature and sunshine hours over Ilorin. But declines with rising relative humidity and rainfall in Ilorin. Though, evaporation piche over Sokoto rises with rising temperature, sunshine hours, solar radiation and wind speed, it declined with rising relative humidity and rainfall. This study shows that relative humidity by far is the most important meteorological factor affecting the level of evaporation piche in the ancient cities of Ilorin and Sokoto.

References

1 Building Nigeria's Response to Climate Change (BNRCC) Report (2011). National Adaptation Strategy and Plan of Action on Climate Change for Nigeria. Prepared for the Federal Ministry of Environment Special Climate Change Unit.

2 Chang, F.G., Chang, L.C., Kao, H.S. & Wu, G.R. (2010). Assessing the effort of meteorological variables for evaporation estimation by self-organizing map neural network. *J. Hydrol.* 384, 118-129.

3 Chattopadhyay, N. & Hulme M. (1997). Evaporation and potential evapotranspiration in India under conditions of recent and future climate change. *Agricultural and Forest Meteorology* 87: 55-73.

4 Cohen, S., Ianetz, A. & Stanhill G. (2002). Evaporative climate changes at Bet Dagan, Isreal, 1964-1998. *Agricultural and Forest Meteorology* 111: 83-91.

5 Hastie T., Tibshirani R. & Friedman J. (2008*). Elements of Statistical Learning, Data Mining, Inference and Prediction,* Second Edition, Springer, Stanford, California.

6 Herch, N.M., Burn, D.H. (2005). *Analysis of trends in evaporation*- phase 1. University of Waterloo, ON Canada.

7 Hess, T.M. (1998). Trends in reference evapotranspiration in the North East Arid Zone in Nigeria. *Journal of Arid Environments* 38: 99-115.

8 Jun, A. & Hideyuk, K. (2004). Pan Evaporation trends in Japan and its relevance to the variability of the hydrologic cycle. *Tenki* 51(9): 667-678.

9 Kay, A.L. & Davies, H.N. (2008). Calculating potential evaporation climate model data. A source of uncertainty for hydrological climate change impacts. *J. Hydrol.*, 358 (3-4), pp. 221-239.

10 Kim, S., Shiri, J., Kisi, O. (2012). Pan evaporation modeling using neural computing approach for different climatic zones. *Water Resour. Manage*. 26(11), 3231-3249.

11 Kisi, O., (2009a). Daily pan evaporation modeling using multi-layer perceptrons and radial basis neural networks. *Hydrol. Process*. 23(2), 213-223.

12 Kisi, O., (2015). Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *J. Hydrol*. 528, 312-320.

13 Obaje N.G. (2009). The Sokoto Basin (Nigerian Sector of the Iullemmeden Basin). In: Geology and Mineral Resources of Nigeria. Lecture Notes in Earth Sciences, vol 120. Springer, Berlin, Heidelberg

14 Peterson, T.C., Golubev, V.S. & Groisman P.Y. (1995). Evaporation losing its strength. *Nature* 377(26): 687-688.

15 ***R*** Core Team, ***R***: A language and environment for statistical computing. ***R*** Foundation for statistical computing, (2018) Vienna, Austria. URL http://www.R-project.org/.

16 Roderick, M.L. & Farquhar, G.D. (2005). Changes in New Zealand pan evaporation since the 1970s. *International Climate of Climatology* 25: 2031-2039.

17 Roderick, M.L., Hobbins, M.T. & Farquhar G.D. (2009). Pan Evaporation trends and the terrestrial water balance. *Geography compass* 3(2): 746-760.

18 Shen, Y.J., Liu, C.M., Liu, M. (2009). Change in pan evaporation over the past 50years in the arid region of China. *Hydrol. Processes* 24, 225-231.

19 Shih, S.F. (1984). Data requirements for evaporation estimation. ASCE, 110(1R3), 263.

20 Wang, G.Q. (2006). Impacts of climate change on hydrology and water resources in the middle reaches of the Yellow River Basin. Hohai University, China.

21 Lunche Wang, Ozgu Kisi, Mohammad Zounemat-Kermani, Li Hui. (2017). Pan Evaporation modeling using six different heuristic computing methods, *Journal of Hydrology*, Volume 544, pp. 407-427.

22 Xie, P. (2009). Spatial-temporal variability and simulation of evapotranspiration in East River Basin. Sun Yat-Sen University, China.

23 Xu, C.Y., Gong, L., Tong, J. & Chen, D. (2006). Decreasing reference evapotranspiration in a warming climate. A case of Changjiang (Yangtze) River catchment during 1970-2000. *Advances in Atmospheric Sciences* 23(4): 513-520.