Hindawi
Journal of Mathematics
Volume 2021, Article ID 9941899, 14 pages
https://doi.org/10.1155/2021/9941899

Research Article

Machine Learning Modelling of the Relationship between Weather and Paddy Yield in Sri Lanka

Piyal Ekanayake[1], **Windhya Rankothge**[2], **Rukmal Weliwatta**[1], **and Jeevani W. Jayasinghe**[1]

[1]*Faculty of Applied Sciences, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka*
[2]*Faculty of Computing, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka*

Correspondence should be addressed to Jeevani W. Jayasinghe; jeevani@wyb.ac.lk

Received 28 March 2021; Revised 27 April 2021; Accepted 15 May 2021; Published 30 May 2021

Academic Editor: Niansheng Tang

This paper presents the development of crop-weather models for the paddy yield in Sri Lanka based on nine weather indices, namely, rainfall, relative humidity (minimum and maximum), temperature (minimum and maximum), wind speed (morning and evening), evaporation, and sunshine hours. The statistics of seven geographical regions, which contribute to about two-thirds of the country's total paddy production, were used for this study. The significance of the weather indices on the paddy yield was explored by employing Random Forest (RF) and the variable importance of each of them was determined. Pearson's correlation and Spearman's correlation were used to identify the behavior of correlation in a positive or negative direction. Further, the pairwise correlation among the weather indices was examined. The results indicate that the minimum relative humidity and the maximum temperature during the paddy cultivation period are the most influential weather indices. Moreover, RF was used to develop a paddy yield prediction model and four more techniques, namely, Power Regression (PR), Multiple Linear Regression (MLR) with stepwise selection, forward (step-up) selection, and backward (step-down) elimination, were used to benchmark the performance of the machine learning technique. Their performances were compared in terms of the Root Mean Squared Error (RMSE), Correlation Coefficient (R), Mean Absolute Error (MAE), and the Mean Absolute Percentage Error (MAPE). As per the results, RF is a reliable and accurate model for the prediction of paddy yield in Sri Lanka, demonstrating a very high R of 0.99 and the least MAPE of 1.4%.

1. Introduction

It is understood that favorable weather conditions as well as other factors like adoption of modern technologies into farming, food preservation techniques, and improved varieties of seeds, fertilizers in cultivation, and so on all contribute to enhanced food security and productivity in the field of agriculture. Among the many progressive steps taken towards the sustainable expansion of major crops grown worldwide, long-term plans for self-sufficiency and raising productivity in paddy cultivation are sensitive issues for agriculture scientists and policymakers because paddy rice continues to be the primary source of food in many countries of the world today and particularly in Asia. With the ever-growing world population towards 10 billion marks by the middle of this century, the demand for rice shall always be on increase and the agriculture technologists will be hard pressed to invent yield-enhancing techniques, as the scope of farming lands for paddy cultivation shall be exhausted within a few years.

Researchers have studied the factors that influence regionwise crop yield differences under technological, biological, and environmental categories [1]. For example, the Random Forest (RF) was used to assess the parameters related to biophysical and socioeconomic environments that affect the growth of paddy [2]. Among the contributory factors mentioned above, it has been found that weather factors account for more on productivity of crops than

others due to their direct and indirect effects [3]. Fred Below ranked seven categorical management factors that impact the corn grain yield and showed that the influence created by the weather on yield is the greatest with 27% contribution compared to other factors like nitrogen, hybrid, and previous crop [4].

Due to this significant influence created by weather on crop yield, it would be a useful exercise to identify the most impactful weather factors and the correlation among them, so that appropriate measures may be contemplated to maximize the effect of conducive factors and minimize that of harmful factors on the paddy yield. Given the uncontrollable and unpredictable nature associated with weather, the researchers' scope is limited to the use of secondary data on regular weather patterns in developing crop-weather models for accurate yield prediction of crops despite occasionally extreme weather conditions.

Some related studies could be found in the literature that had used the following regression techniques to address the above topic in some other countries. Sharma and Joshi examined the spatial and temporal performance of rice production and yield and the factors determining the acreage and yield of paddy in coastal regions of India [5]. They used the ordinary least squares to estimate the equations and fitted multiple regressions to interdistrict data for the period from 1984/85 to 1988/89 to find out the extent to which the variables, including irrigation, fertilizer use, rainfall, and area under high yield varieties, are responsible for the growth of the paddy yield. It was found that rainfall and fertilizer use are the most important factors associated with positive coefficients, to increase the yield. A crop-weather model was used for the prediction of paddy yield in Tamil Nadu, India, using a full model and stepwise regression analysis [6]. This study, having subjected seven variables from 10 years of data into stepwise regression, predicted the paddy yield of one paddy growing season with a coefficient of determination (R^2) of 0.9234 using only four predictors, namely, percentage of rice area, number of days with minimum temperature, average daily minimum temperature, and monthly average solar radiation. In this paper, Power Regression (PR) and three Multiple Linear Regression (MLR) models with stepwise selection, forward selection, and backward elimination of variables are used to relate the paddy yield to weather indices and their performance shall be compared with that of the more powerful nonparametric methods of PR and RF to identify the most suitable model(s) in the Sri Lankan context characterized by two major paddy growing seasons in nine regions with different weather conditions.

Machine learning techniques have also been used to develop crop-weather models and to understand the most influential weather factors. Konduri et al. compared the performance of linear and nonlinear regression models in terms of R^2 and the Root Mean Square Error (RMSE) and found that Support Vector Regression (SVR) and RF are capable of producing comparatively better performance over the linear models of Principle Component Regression and Ridge Regression in assessing the impact of climate on the crop yield [7]. They further highlighted the accuracy of RF regression while attributing its superiority in handling data to multicollinearity and extracting nonlinear interactions. A comparative assessment had been conducted on the linear regression and two versions of RF for extracting the relative importance of the regressor variables [8]. As reported in this study [8], linear regression would collapse when there are more variables than observations, whereas being a nonparametric method, RF emerged to be more robust to explain nonlinearities and interactions known to exist between weather indices and crop yield. Shi and Horvath had also shown that RF dissimilarity could deal with mixed variable types (categorical and ordered) in a straightforward manner and that it was consistent with respect to routine transformations of the variable values and strong to outliers [9]. Due to the reported superiority of RF in developing crop-weather models, it was also used in this research to develop a paddy-weather model for Sri Lanka.

Although the weather factors were known to control the crop yield to a greater extent, a comprehensive study focusing on their relative importance and correlation with the paddy yield has not yet been conducted to explore the situation in Sri Lanka. Therefore, the objectives of the present study were focused on investigating the most impactful weather indices on paddy yield in Sri Lanka. In light of numerous modelling techniques cited above, it was possible to narrow down the choice of methods that would help achieve the objectives of this study. Due to the overwhelming success reported in using RF, it will be used to shed more light on interregressor correlation, which is an important determinant of the behavior of variable importance matrix.

In Section 2 of this paper, the models, methodology, and the scope of the data analysis shall be described. The research findings are discussed in detail in Section 3 with reference to variable importance, correlation, and regression models, followed by the validation of results based on observed and predicted yields. Section 4 carries the summary of the conclusions drawn from the study for the Sri Lankan context.

2. Materials and Methods

2.1. Data. Eleven years of secondary data on paddy yield were obtained from the reports published by the Department of Census and Statistics, the premier state institute in Sri Lanka, maintaining the official repository of information on diverse fields collected using appropriate scientific methods and instruments. The temporal scope of data included the two main paddy cultivation seasons spanning from May to August (Yala season) and September to March (Maha season) of the ensuing year during the period from 2009 to 2019, while the spatial coverage encompassed seven administrative districts, which together contribute to nearly 62% of the overall annual paddy production in Sri Lanka (Figure 1).

Table 1 presents the areawise (districtwise) average percentages contributing to the overall annual paddy production of Sri Lanka, which is about 2.7 million tons and satisfies about 95% of the domestic requirement. Paddy is

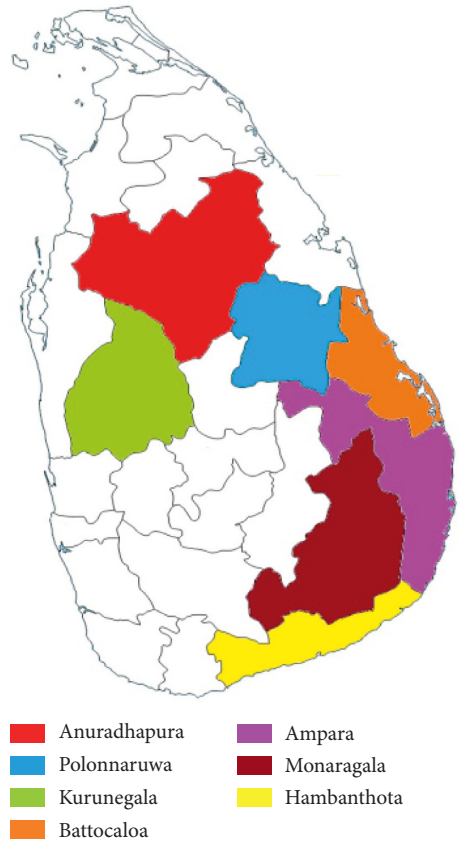

Figure 1: Study areas.

Table 1: Paddy yield in the study areas.

District	Average contribution to the paddy production in Sri Lanka (%)	Season	Yield (t/ha)		
			Mean	Median	Range
Ampara	15.57	Yala	4.8	4.8	1.1
		Maha	4.6	4.8	2.0
Polonnaruwa	10.64	Yala	4.9	5.0	1.3
		Maha	5.0	5.0	1.6
Kurunegala	10.56	Yala	3.7	3.6	0.6
		Maha	4.2	4.0	1.5
Anuradhapura	9.07	Yala	4.6	4.5	1.4
		Maha	4.7	4.6	1.7
Batticaloa	6.17	Yala	5.2	5.1	1.6
		Maha	5.8	5.8	1.4
Hambantota	6.23	Yala	4.2	4.1	1.8
		Maha	3.1	3.1	1.9
Monaragala	3.49	Yala	4.0	4.0	1.4
		Maha	4.2	4.2	0.9
Total	61.73				

cultivated by about 1.8 million farming families spreading across the country in an estimated extent of 870,000 ha annually. It can be traced from the table that the mean yield of the seven districts during the Yala season is in the range of 3.7 to 5.2 t/ha and during the Maha season varies within a slightly wider range of 3.1 to 5.8 t/ha. Except in Ampara and Hambantota districts, the mean yield during Maha season is generally higher than that during Yala season. It can also be noted that the most fertile yields are produced by Batticaloa and Polonnaruwa districts in both seasons.

Weather data were purchased from another state institute, the Department of Meteorology in Sri Lanka, for the same period as for the paddy yield data. The total rainfall during a cultivation season was used with the seasonal

averages of eight more monthly mean weather indices in relative humidity (minimum and maximum), temperature (minimum and maximum), wind speed (morning and evening), evaporation, and sunshine hours. Thus, the above temporal and spatial extent provided a total of 11 years × 7 districts × 2 seasons of data for the analysis carried out using MLR, PR, and RF. In MLR, three types of variable selection methods, namely, stepwise, forward selection, and backward elimination, were employed.

Table 2 summarizes the amount of total rainfall received during the period of cultivation and the means of the other weather indices in the seven geographical regions covered by the data. It can be noted that the highest rainfall during the paddy growing seasons is recorded at Batticaloa district, followed by Polonnaruwa district and the lowest rainfall has occurred at Hambantota district. The least minimum relative humidity prevails at Polonnaruwa and Monaragala districts, while the highest maximum relative humidity prevails at Kurunegala, Anuradhapura, and Batticaloa districts. The minimum temperature has fallen to about 22°C at Polonnaruwa and Monaragala districts and the maximum has gone up to 33.5°C at Polonnaruwa district. The highest evaporation recorded at Polonnaruwa district is consistent with the most sunshine hours compared to other districts. The morning wind speed is the strongest (5.8 km/h) at Anuradhapura district in the North-Central province, followed by Hambantota district with 4.8 km/h in the Southern province of Sri Lanka, while the weakest is reported at Kurunegala, Batticaloa, and Monaragala districts. Though weaker in the morning, Batticaloa on the eastern coast records the strongest evening winds (6.9 km/h), followed by Anuradhapura district. In general, it may be inferred that a very windy environment prevails at Anuradhapura district rich with many large reservoirs, while Kurunegala and Monaragala remain relatively tranquil compared to other districts. Further, the evening winds on average are stronger than the morning winds in all districts.

2.2. Variable Importance. The relative importance of predictors is usually measured by evaluating how much each predictor contributes to increasing the model accuracy [10]. Therefore, the variable importance (or feature importance) techniques refer to a set of techniques, which assign scores to predictors and indicate the relative importance of each predictor when making an accurate prediction. It provides an insight into the dataset as well as to the predictive model and is useful for the improvement of the predictive model. Further, it highlights the most significant predictors and the least significant predictors [10]. Therefore, it could be used as the basis for gathering more or different data for the model. Based on the significance of each predictor, a feature selection can be performed to retain only the most significant predictors in the prediction model. It simplifies the problem being modelled and speeds up the modelling process, thus improving the overall performance of the model.

In this research, the in-built variable importance method of RF regression model [11, 12] was used to understand how much each predictor (weather index) contributes towards the yield prediction. The RF regression first generates a set of decision tree models that use diverse combinations of predictors. Each decision tree is a set of internal nodes and leaves grown on a bootstrap sample of the original dataset. Only a random subset of the predictors is considered as splitting candidates at each split in the trees. Splitting rules in RF regression maximize the decrease of the impurity introduced by a split. RF regression measures how each predictor decreases the impurity of the split and the predictors with the highest decrease are selected for the internal node. For all trees and each predictor, an average value on how it decreases the impurity is calculated and it is considered as the measure of the variable importance for that predictor [11, 12].

For each decision tree, RF regression calculates nodes' importance using Gini Importance, assuming only two child nodes (binary tree). The importance of node j is defined as

$$ni_j = w_j C_j - w_{\text{left}(j)} C_{\text{left}(j)} - w_{\text{right}(j)} C_{\text{right}(j)}, \tag{1}$$

where w_j is the weighted number of samples reaching node j, C_j is the impurity value of node j, left (j) is the child node from left split on node j, and right (j) is the child node from right split on node j. The importance of each feature i on a decision tree is then calculated as

$$fi_i = \frac{\sum j\text{: node } j \text{ splits on feature } i \cdot ni_j}{\sum k \in \text{all modes} \cdot ni_k}. \tag{2}$$

Next, the feature importance values are normalized and the normalized feature importance for i in tree j is specified as

$$\text{norm } fi_i = \frac{fi_j}{\sum j \in \text{all features} \cdot fi_j}. \tag{3}$$

The final feature importance at the RF level is its average over the total number of trees (T).

$$\text{RF } fi_i = \frac{\sum j \in \text{all trees} \cdot \text{norm } fi_{ij}}{T}. \tag{4}$$

2.3. Pearson's Correlation Coefficient (R). The correlation between the yield and each weather index was determined to quantify its impact and also to identify whether the impact is positive or negative. Pearson's correlation coefficient and Spearman's correlation coefficient were calculated using the programming language R studio (version 1.3.1093). Pearson's correlation coefficient is a test statistic that measures both the strength and direction of a pairwise linear relationship between two quantitative continuous variables [13]. It is calculated based on the following formula:

$$R = \frac{\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \overline{x})^2 \sum_{i=1}^{N} (y_i - \overline{y})^2}}, \tag{5}$$

where, in this study, x_i and y_i are the observations of a pair of variables from the yield and the weather indices mentioned in Section 2.1. $\overline{x}$ and $\overline{y}$ are the means of the two variables.

Table 2: Mean weather in the study areas during the period of paddy cultivation.

District	Mean weather								
	Rainfall (mm)	Minimum relative humidity (%)	Maximum relative humidity (%)	Minimum temperature (°C)	Maximum temperature (°C)	Evaporation (mm)	Sunshine hours	Morning wind speed (km/h)	Evening wind speed (km/h)
Ampara	741.5	71.7	76.0	24.6	33.0	3.7	7.1	3.4	5.5
Polonnaruwa	896.3	60.8	74.2	21.9	33.5	4.4	7.6	3.6	3.9
Kurunegala	644.2	72.0	83.4	23.3	32.3	3.1	6.8	2.9	3.9
Anuradhapura	681.5	70.2	83.2	23.6	32.2	3.6	7.4	5.8	6.1
Batticaloa	994.8	71.0	83.4	25.4	32.2	3.6	7.3	3.0	6.9
Hambantota	574.3	73.0	78.0	24.1	32.6	4.2	6.4	4.8	5.2
Monaragala	801.7	64.1	78.0	22.3	33.0	3.2	6.3	2.9	4.2

A positive correlation coefficient implies an increase of both variables in the same direction and a negative value means the change of variables in opposite directions. The correlation matrix thus obtained is given in Table 3. Further, nonzero values of R close to ±1 are the evidence for strong linear associations between the variables, and values close to zero indicate no such relationship. Pearson's correlation is appropriate for linearly related variables, each of which has a normal (Gaussian, "bell-shaped curve," parametric) distribution, while Spearman's rank correlation can be used on nonlinearly related, nonnormal distributions (nonparametric) [14].

2.4. Spearman's Correlation (R_s). As some studies had reportedly shown nonlinear relationships between the yield and weather indices [7], it was decided to examine the pairwise Spearman's correlation coefficient within the paddy yield and the same weather indices paired exhaustively, as summarized in Table 4. It can vary within the range from −1 to +1, such that the limits imply a perfect monotonic relationship [15], and it is given as follows:

$$R_s = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N\left(N^2 - 1\right)}, \tag{6}$$

where d_i is the difference between the two ranks of each observation and N is the number of observations.

A value of R_s close to +1 indicates a strong positive association of ranks, −1 indicates a strong negative association of ranks, and zero indicates a weaker or no association between the ranks. A nonlinear relationship may be present even if this coefficient is zero. One of the advantages is that Spearman's correlation coefficient could be used when the assumptions for Pearson's correlation coefficient, namely, normality, linearity, and the continuous nature of variables, are no longer valid.

2.5. Multiple Linear Regression. As the number of observations is much more than the number of variables, linear regression is known to be a strong classical parametric method [8]. In this study, MLR was used to examine how the independent variables are related to the dependent variable. Once the relation between the dependent variable and independent variables is identified, it can be used to make more powerful and accurate predictions on the dependent variable. The paddy yield was taken as the dependent variable, while the nine weather indices of the corresponding seasons were used as independent variables. Being an extension of the ordinary least squares regression, the yield in MLR is expressed as follows:

$$\begin{aligned}\text{yield} = {} & \beta_0 + \beta_1 T_{\min} + \beta_2 T_{\max} + \beta_3 SH + \beta_4 E + \beta_5 R \\ & + \beta_6 H_{\min} + \beta_7 H_{\max} + \beta_8 W_m + \beta_9 W_e + \varepsilon,\end{aligned} \tag{7}$$

where β_0 is the intercept (a constant), β_1 to β_9 are the regression coefficients of the input variables, and ε is the random error under the assumption that it is normally distributed with mean zero and constant variance.

Three MLR methods differed according to the selection procedure of variables, namely, forward (step-up) selection, backward (step-down) elimination, and the stepwise selection, which were used. The stepwise regression is a combination of the other two techniques wherein variables are added stepwise after verifying their significance against a tolerance level. In the forward (step-up) selection method, the predictor variables (weather indices) are added in the decreasing order of their correlation with the dependent variable (yield). An opposite process takes place in the backward (step-down) elimination method in which each predictor variable not contributing to the regression equation is removed.

2.6. Power Regression. PR is a nonlinear regression model in which the output is modelled in proportion to the power of the explanatory variables. In PR, the function is a power (polynomial) equation of the form $y = ax^b$, where x has to be nonzero. The equation predicts y-values lying within the plotted values of x, as it is less reliable to predict y-values that lie outside the plotted values. In this research, the paddy yield of Yala or Maha season in any year was taken as the dependent variable, while the corresponding weather indices were used as independent variables. It can be expressed as follows:

$$\text{yield} = aT_{\min}^{b} T_{\max}^{c} SH^{d} E^{e} R^{f} H_{\min}^{g} H_{\max}^{i} W_m^{j} W_e^{k}, \tag{8}$$

where $a, b, c, \ldots, k$ are constants.

Table 3: Pearson's correlation matrix.

	Yield	H_{min}	T_{max}	W_e	E	W_m	R	SH	T_{min}	H_{max}
Yield	1.00									
H_{min}	−0.05	1.00								
T_{max}	0.21	−0.79	1.00							
W_e	−0.20	0.15	−0.07	1.00						
E	0.39	−0.56	0.78	0.75	1.00					
W_m	0.22	0.03	−0.07	0.67	0.68	1.00				
R	−0.14	0.57	−0.79	−0.22	−0.59	−0.15	1.00			
SH	0.07	0.51	0.58	0.47	0.70	0.51	−0.52	1.00		
T_{min}	−0.17	0.11	0.27	0.52	0.33	0.17	−0.46	0.22	1.00	
H_{max}	−0.30	0.84	−0.81	0.60	−0.87	0.71	0.54	−0.93	0.09	1.00

2.7. Random Forest. RF is a widely used supervised learning-based machine learning technique that has proved its efficiency in modelling the crop yield owing to its sound performance in many prediction domains [16, 17]. In this research, RF regression method was employed as it had been successfully used in agriculture applications such as predicting the yield of different crops (wheat, maize, and potato) accurately with climate and biophysical predictors at global and regional scales [18]. Also, its nonlinear nature is helpful when developing a reliable model to understand the relationships among the climate, biophysical predictors, and the yield [11].

RF constructs a predictive model and estimates the relative importance of predictors [12]. It first generates a set of decision tree models that use diverse combinations of predictors and thresholds to explain datasets, which are generated for the individual trees by sampling from original data. Then, it takes an overall average of these tree model outputs as a prediction, which is known as ensemble modelling. Instead of just averaging the prediction of trees, RF uses two key concepts that give it the name random: (1) random sampling of training observations when building trees and (2) random subsets of features for splitting nodes [11]. RF builds multiple decision trees and merges their predictions together to get a more accurate and stable prediction rather than relying on individual decision trees. The intrinsic variable selection facilitates the dissimilarity of RF to handle a large number of variables [9]. The relative importance of predictors is usually measured by evaluating how much each predictor contributes to increasing the model accuracy [12].

In this research, first, the data were feature normalized as an input set X: {Rainfall, Minimum relative humidity, Maximum relative humidity, Minimum temperature, Maximum temperature, Evaporation}, and an output set Y: {Predicted yield}. Then the data were split into a training set and a testing set, comprised of 80% and 20%, respectively, to fit the RF on the input data. Next, the data were fetched into the RF model with 10 decision trees where the depth of each tree was 5 levels. Finally, the accuracy of the model was evaluated in terms of some statistical parameters.

2.8. Evaluation of the Models. After developing the models of RF, MLR with stepwise selection, MLR with forward (step-up) selection, MLR with backward (step-down) elimination, and PR, their performance was evaluated in terms of the correlation coefficient (R), RMSE, Mean Absolute Error (MAE), and the Mean Absolute Percentage Error (MAPE).

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}}, \\ \text{MAE} &= \frac{1}{N}\sum_{i=1}^{N}\left|x_i - y_i\right|, \\ \text{MAPE} &= \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i - y_i}{x_i}\right| \times 100, \end{aligned} \tag{9}$$

where x and y are the actual and estimated yields, respectively, and N is the number of observations. The lower the RMSE, MAE, and MAPE and the closer the R to 1, the more the accurate models that fit actual paddy yield well with predicted yield.

3. Results and Discussion

The feature importance of each independent variable on the paddy yield was measured as a fraction and the distribution of the two most important variables was examined to clarify their correlation values with the paddy yield in correlation matrices. The correlation of each weather index with the yield and the remaining weather indices was quantified using Pearson's correlation method and Spearman's correlation method. Strong and moderate correlations were distinguished from the weaker correlations based on three ranges. The performance of the five models can be understood in comparison with each other in terms of the statistical measures of R, RMSE, MAE, and MAPE. The distribution of errors of the predicted yield arising from the MLR (Stepwise), PR, and RF methods was also illustrated.

3.1. Variable Importance and Correlation. Minimum relative humidity was found to be the most important independent variable (Figure 2). However, neither Pearson's Correlation Matrix nor Spearman's Correlation Matrix indicated a higher correlation between the minimum relative humidity and paddy yield (Tables 3 and 4). Correlation between the independent variable (minimum relative humidity) and the

Table 4: Spearman's correlation matrix.

	Yield	H_{min}	T_{max}	W_e	E	W_m	R	SH	T_{min}	H_{max}
Yield	1.00									
H_{min}	0.01	1.00								
T_{max}	0.18	−0.79	1.00							
W_e	−0.10	0.17	−0.09	1.00						
E	0.42	−0.44	0.79	0.63	1.00					
W_m	0.25	0.21	−0.18	0.63	0.49	1.00				
R	−0.12	0.62	−0.84	−0.24	−0.67	−0.01	1.00			
SH	0.03	−0.46	0.57	0.38	0.59	0.36	−0.53	1.00		
T_{min}	−0.13	0.04	0.32	0.53	0.47	0.11	−0.56	0.25	1.00	
H_{max}	−0.30	0.80	−0.84	0.64	−0.68	0.80	0.66	−0.80	−0.07	1.00

dependent variable (paddy yield) was investigated to understand this incoherence. It was observed that the relationship was not identified in terms of Pearson's correlation due to its nonlinear behavior (Figure 3(a)). In order to check the reason resulting in a less Spearman's correlation, the distribution of minimum relative humidity data was plotted, a histogram was generated, and the frequency density curve was superimposed on it. It was observed that the distribution of data is not normal (Figure 3(b)). Particularly, the nonmonotonic behavior in the relationship between the minimum relative humidity and paddy yield disturbs, identifying Spearman's correlation (Figure 3(a)).

The second most important independent variable is the maximum temperature. Both Pearson's correlation and Spearman's correlation indicate a positive relationship between the maximum temperature and the paddy yield. The positive Pearson's correlation is coherent with the linear relationship (Figure 4(a)). Similarly, it exhibits a nonlinear relationship, which is again positive resulting in a positive Spearman's correlation value (Figure 4(b)). As the optimum temperature at all the growth stages of rice, that is, from emergence to ripening and harvesting and particularly for flowering in rice plant, ranges from 27°C to 32°C [19], no increment in the paddy yield is shown above the temperature of 32°C. The distribution of maximum temperature data was also investigated and found normal (Figure 4(c)). The dependent variable and paddy yield also demonstrated a normal distribution resulting in a considerable correlation between the two indices (Figure 4(d)).

Wind speed is the third most important variable, whereas the winds in the morning and evening affect the yield contrarily such that wind in the morning is showing a positive correlation with the yield and in the evening is correlating negatively. This contrasting correlation of winds may be due to the negative effect caused by stronger evening winds (Table 2). It is reported in literature too that strong winds during the flowering stage hinder the fertilization in paddy [20]. Evaporation correlates positively to the paddy yield, while the rainfall correlates negatively. The importance as well as the correlation of the other two variables, namely, the number of sunshine hours and minimum temperature, is minimal.

Strong correlations were identified if both correlation values between two indices are within the interval [0.75, 1.0] or [−0.75, −1] and mediocre correlations if at least one of the values is within the interval [0.50, 0.74] or [−0.50, −0.74] and the other value lies in the higher (strong) interval. Accordingly, both strong and mediocre correlations of positively and negatively associated variables are summarized in Table 5.

3.2. Regression Models. A total of five crop-weather models were developed in this study taking both linear and nonlinear aspects into consideration and their performance is summarized in Table 6. Based on the performance indicators, it can be comprehended that there is little difference between the MLR methods with forward selection and backward elimination, as the corresponding statistical measures are very close to each other. Comparatively, the MLR method with stepwise regression and the nonlinear PR method have shown similar and better performance substantiated by the statistical performance indicators.

The regression equations emerged from stepwise MLR and PR which are given in (9) and (10), respectively, wherein the former model is represented in terms of five weather indices. The PR model retained the morning wind speed instead of the evening wind speed. Moreover, the similarity of these two models is further evident from their identical error distributions depicted in Figures 5(a) and 5(b).

$$Y_{MLR} = 11097 - 0.55RF + 36RH_{min} - 179.9T_{max} - 144.2T_{min} + 81.6\text{wind}_e, \quad (10)$$

$$Y_{PR} = 1244030RF^{-0.07}RH_{min}^{0.46}T_{max}^{-1.2}T_{min}^{-0.98}\text{wind}_m^{0.08}. \quad (11)$$

The most encouraging results were generated by the nonlinear RF method with the highest correlation coefficient and the least RMSE, MAE, and MAPE (Table 6). The higher correlation is coherent with the excellent coincidence of the yield predicted by the model with the actual yield, as shown in Figure 5(c). The superiority of the RF-based results can be observed in Figure 6 too, which shows the distribution of the percentage of data samples against six consecutive intervals of error. Errors of the stepwise MLR model and the PR model are of comparable magnitude and distributed over the error intervals, while 40% and 60% of data samples have errors less than 1% and within 1–5%, respectively, for the RF model. The variation of predicted paddy yield against the

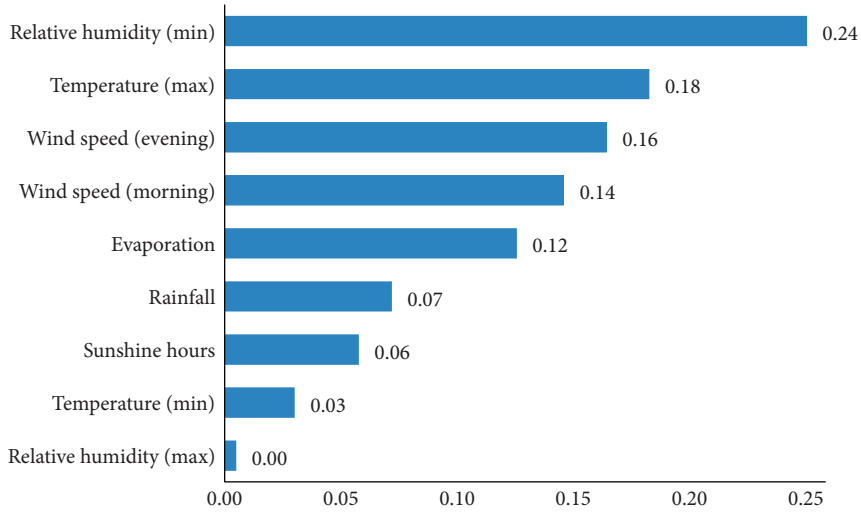

Figure 2: Variable importance.

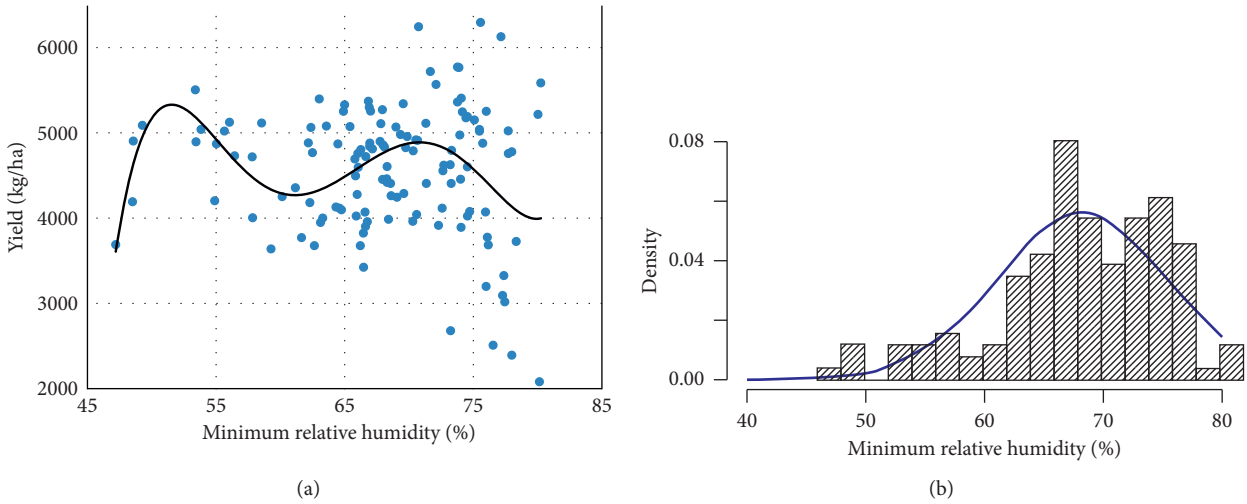

Figure 3: Distribution of minimum relative humidity data: (a) relationship with paddy yield and (b) data distribution.

actual yield of the RF model is illustrated in Figure 7. It also indicates that all the predicted yield values are very close to the corresponding actual yield.

3.3. Discussion. Researchers have used numerous statistical and machine learning techniques to develop crop-weather models for a variety of crops such as paddy, wheat, and corn. A summary of relevant research studies is presented in Table 7. In these studies, different weather indices were suggested as the most influential independent variable(s). The reason behind diverse conclusions is the differences in the weather at the study areas, which varies over a wide range. For example, temperature less than 19°C is critical for inducing grain sterility in paddy [27] but the temperature in the equatorial paddy growing areas does not usually drop down that much. Similarly, the optimum relative humidity for paddy cultivation lies between 60% and 80%, while values higher than 85% are critical [28]. However, the spikelet fertility was not always inhibited only by high relative humidity [29]. Rather, it induces almost complete paddy sterility at a temperature of about 35°C [27]. Hence, higher temperatures with high relative humidity decrease paddy yield [30] proving that the combined effect of temperature and relative humidity is a predominant factor in paddy cultivation [28]. In this sense, a comprehensive analysis in the area of interest is required to understand the relationship between weather and paddy yield there.

In the context of Sri Lanka where rice is the staple food, the effects of climatic variation were extensively researched [31–34]. However, in most of the research studies, only a few climatic factors were considered. Therefore, the readers, particularly the responsible authorities, are not given a clear picture of the influence created by weather indices on the paddy yield. In this research, the correlation between the paddy yield and all the related meteorological factors is

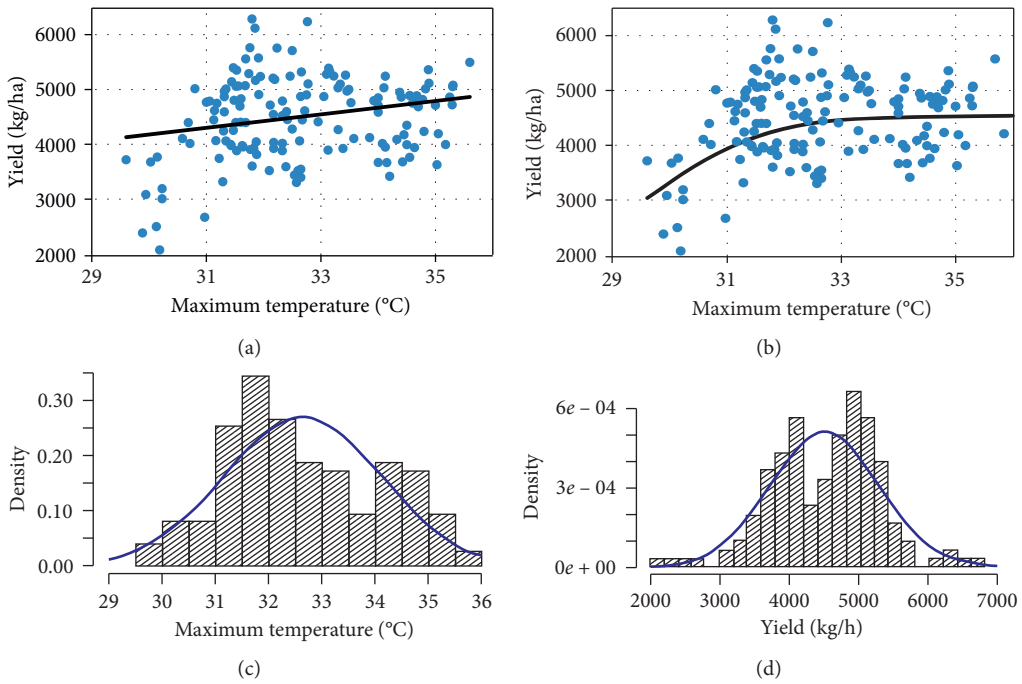

Figure 4: Distribution of maximum temperature data: (a) linear relationship between the maximum temperature and the paddy yield; (b) nonlinear relationship between the maximum temperature and the paddy yield; (c) normal distribution of maximum temperature data; (d) normal distribution of the dependent variable.

Table 5: Correlation between the weather indices.

Level of correlation	Positively correlated pairs of weather indices	Negatively correlated pairs of weather indices
Strong	Maximum relative humidity and minimum relative humidity, evaporation, and maximum temperature	Maximum temperature and minimum relative humidity, rainfall and maximum temperature, maximum relative humidity and maximum temperature, maximum relative humidity, and sunshine hours
Mediocre	Rainfall and minimum relative humidity, sunshine hours and maximum temperature, evaporation and evening wind, sunshine hours and evaporation, maximum relative humidity and morning wind, and maximum relative humidity and rainfall	Maximum relative humidity and evening wind, rainfall and evaporation, maximum relative humidity and evaporation, sunshine hours, and rainfall

quantified and the importance of each factor is identified. This research can be extended to study the influence of weather indices at different stages of paddy cultivation by using weekly weather data. Further, the most influential nonclimatic factors may be identified and their influence can be investigated. These findings will be useful for the agriculture authorities and policymakers to ponder appropriate measures for increasing the paddy yield by mitigating negative effects and optimizing the positive effects through crop management.

Though paddy yield prediction models were developed by applying numerous techniques [35, 36], this is the first research study on developing a crop-weather model for the paddy yield in Sri Lanka. This research can be extended for the prediction of paddy yield for future seasons or years if the independent variables are available as projected climatic variables. When the future weather conditions are estimated or forecast, they can be applied to the models developed in this research for predicting the future paddy yield. Projecting future climate under different scenarios (e.g., Representative

Table 6: Performance of the regression models.

Technique	RMSE	R	MAE	MAPE (%)
MLR: forward method	489	0.54	374	9.2
MLR: backward method	483	0.53	374	9.2
MLR: stepwise method	472	0.75	361	8.9
PR	485	0.75	356	8.7
RF	71	0.99	60	1.4

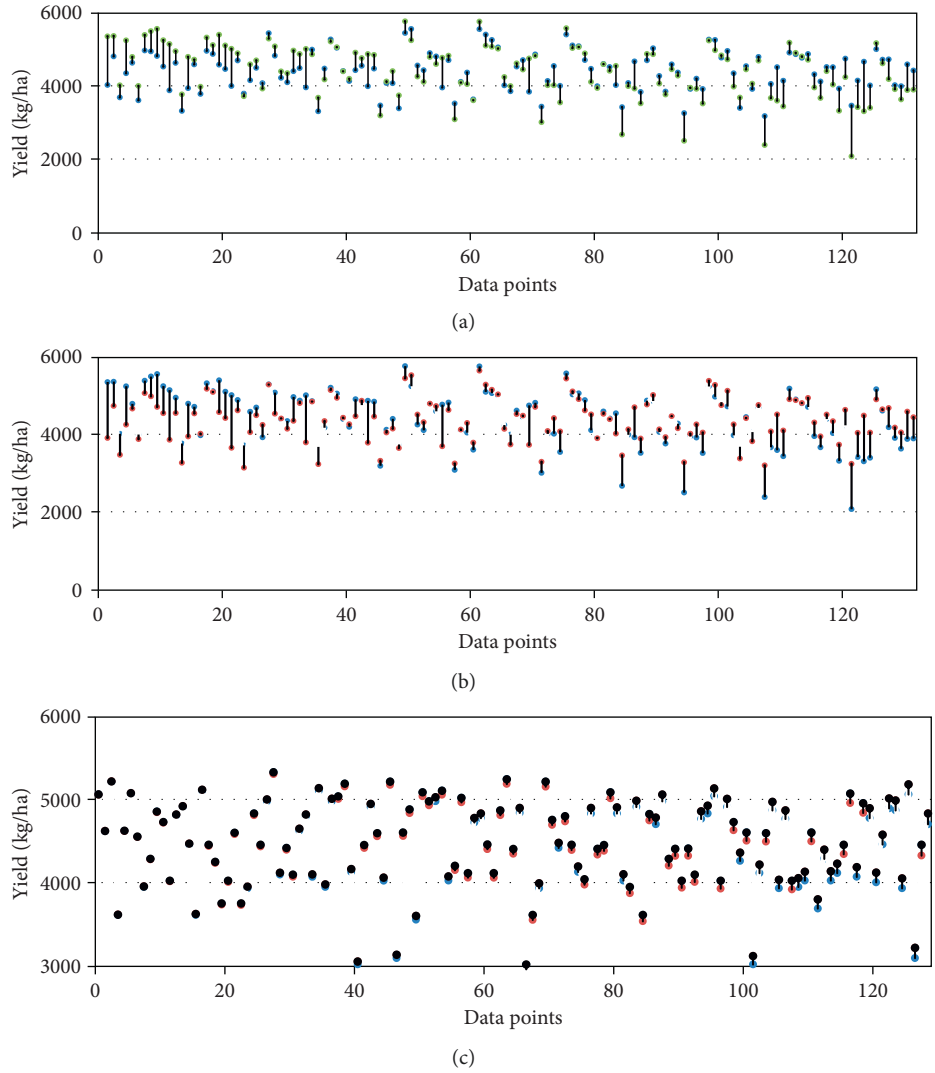

Figure 5: Error of the yield predicted by applying regression techniques: (a) MLR (stepwise), (b) PR, and (c) RF.

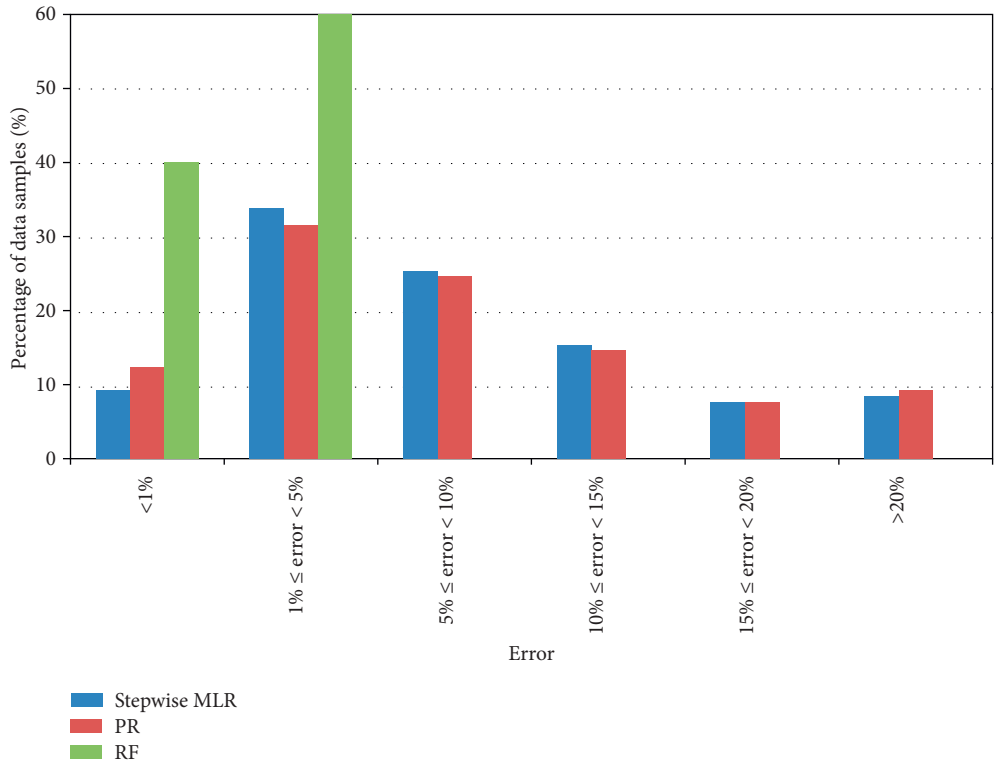

Figure 6: Distribution of error of the predicted yield.

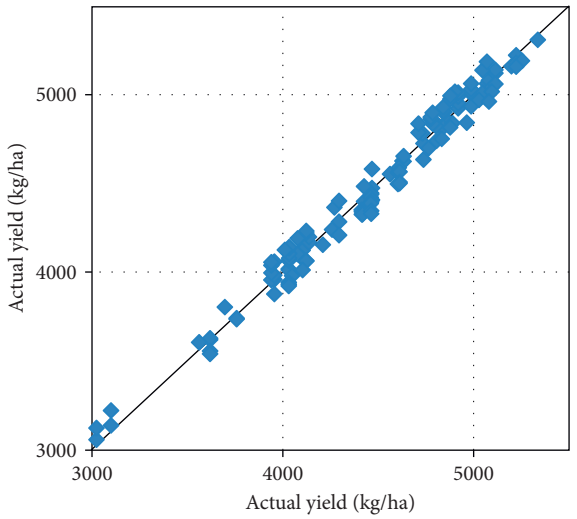

Figure 7: Actual versus predicted yield of the RF model.

Table 7: Crop-weather models.

Ref.	Crop	Country	Evaluation criteria	Weather indices	The most influential weather indices
[6]	Paddy	India	Full model and stepwise MLR	Rice area, number of days with minimum temperature below 22°C, average daily temperature (maximum and minimum), sunshine hours, rainfall, and solar radiation	Solar radiation
[21]	Corn	USA	Kincer's method	Precipitation, temperature, sunshine, and relative humidity	Relative humidity
[22]	Crops	Uganda	MLR	Precipitation, temperature, and CO_2 emissions	CO_2 emissions
[23]	7 crops including paddy and corn	Taiwan	MLR	Temperature and precipitation	Temperature and precipitation
[24]	Paddy	India	Gaussian process regression (GPR) and lasso regression	Temperature, average humidity, rainfall, wind speed, UV index, sun hours, and pressure values	Rainfall
[25]	Wheat	China	RF, SVM, and GPR	Maximum temperature, minimum temperature, drought index, and precipitation	Minimum temperature
[26]	Paddy	Korea	Random forest	Temperature (minimum, mean, and maximum) and sunshine hours	Minimum temperature and sunshine hours

Concentration Pathway) is widely reported [37–39] and one such climate projection scenario can be applied in a future research. As the correlation coefficient of the RF model applied here is 0.99 with very low MAPE of 1.4%, it can be used as a highly accurate yield prediction model.

4. Conclusions

This study was carried out with data available at the Department of Meteorology and the Department of Census and Statistics of Sri Lanka with the objective of extracting the most influential weather factors on the paddy yield in Sri Lanka. The data covered seven major paddy growing regions that account for nearly two-thirds of the overall country production over eleven years in both agricultural seasons. A total of five regression techniques that can model linear relationships as well as nonlinearities and interactions were used. Of these, the RF model was the most accurate regression method. The difference in performance between the forward selection and backward elimination methods of the MLR was insignificant, while the stepwise MLR method was better and remained on par with the PR method. However, the excellence and the accuracy of the RF model were evidently proved by the statistical performance indicators as well as the distribution of errors between the actual yield and model produced yield. This research study may be extended by applying projected climate conditions on the RF model for the prediction of future paddy yield. The ability to predict the future yield will be beneficial to the agriculture authorities to ensure food security. Such projections are useful at macrolevel as the country's economic activities are dominated by the agriculture sector in which the major crop is paddy.

RF regression was used to rank the weather indices affecting the paddy yield in Sri Lanka. The minimum relative humidity emerged as the most impactful weather index having a nonlinear correlation with the paddy yield, followed by maximum temperature which showed both linear and nonlinear relationships with the paddy yield. The morning wind speed was proved to be positively correlated, while the evening wind was negatively correlated with the paddy yield. Pearson's and Spearman's correlation matrices provided further insight into the degree of association between the pairwise weather indices. The weather indices of maximum and minimum relative humidity and evaporation with maximum temperature showed strong positive correlations. Nevertheless, maximum temperature, rainfall, and maximum relative humidity were negatively correlated with humidity, maximum temperature, and sunshine hours, respectively. In future research studies, nonclimatic factors may also be incorporated and their importance may be investigated.

Data Availability

The data used for the research are available from the corresponding author upon request, subject to the approval of the relevant authorities.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are grateful to the Department of Census and Statistics and the Department of Meteorology, Sri Lanka, for providing past records of paddy harvest, yield, and climate data.

References

[1] T. N. Liliane and M. S. Charles, "Factors affecting yield of crops," *Agronomy-Climate Change & Food Security*, vol. 9, 2020.

[2] J. F. Djagba, L. O. Sintondji, A. M. Kouyaté et al., "Predictors determining the potential of inland valleys for rice production

development in West Africa," *Applied Geography*, vol. 96, pp. 86–97, 2018.

[3] K. R. Paltasingh and P. Goyari, "Statistical modeling of crop-weather relationship in India: a survey on evolutionary trend of methodologies," *Asian Journal of Agriculture and Development*, vol. 15, pp. 43–60, 2018.

[4] F. E. Below, "The seven wonders of the corn yield world," in *Proceedings of the 2008 Illinois Crop Protection Technology Conference*, p. 86, Urbana, IL, USA, January 2008.

[5] V. P. Sharma and P. K. Joshi, "Performance of rice production and factors affecting acreage under rice in coastal regions of India," *Indian Journal of Agricultural Economics*, vol. 50, pp. 153–167, 1995.

[6] K. Kandiannan, R. Karthikeyan, R. Krishnan, C. Kailasam, and T. N. Balasubramanian, "A CropWeather model for prediction of rice (*Oryza sativa* L.) yield using an empirical-statistical technique," *Journal of Agronomy and Crop Science*, vol. 188, no. 1, pp. 59–62, 2002.

[7] V. S. Konduri, T. J. Vandal, S. Ganguly, and A. R. Ganguly, "Data science for weather impacts on crop yield," *Frontiers in Sustainable Food Systems*, vol. 4, p. 52, 2020.

[8] U. Grömping, "Variable importance assessment in regression: linear regression versus random forest," *The American Statistician*, vol. 63, no. 4, pp. 308–319, 2009.

[9] T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," *Journal of Computational and Graphical Statistics*, vol. 15, no. 1, pp. 118–138, 2006.

[10] J. Rogers and S. Gunn, "Identifying feature relevance using a random forest," in *Subspace, Latent Structure and Feature Selection. SLSFS 2005. Lecture Notes in Computer Science*, C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, Eds., Vol. 3940, Springer, Berlin, Germany, 2006.

[11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science & Control Engineering*, vol. 2, no. 1, pp. 602–609, 2014.

[13] A. Ly, M. Marsman, and E.-J. Wagenmakers, "Analytic posteriors for Pearson's correlation coefficient," *Statistica Neerlandica*, vol. 72, no. 1, pp. 4–13, 2018.

[14] J. Hauke and T. Kossowski, "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data," *Quageo*, vol. 30, no. 2, pp. 87–93, 2011.

[15] A. K. Sharma, *Text Book of Correlations and Regression*, Discovery Publishing House, New Delhi, India, 2005.

[16] Y. Guo, H. Xiang, Z. Li, F. Ma, and C. Du, "Prediction of rice yield in East China based on climate and agronomic traits data using artificial neural networks and partial least squares regression," *Agronomy*, vol. 11, no. 2, 2021.

[17] P. S. T. Muthusinghe, W. Wand, A. M. H. Saranga et al., "Towards smart farming: accurate prediction of paddy harvest and rice demand," in *Proceedings of the IEEE R10 HTC 2018*, Colombo, Srilanka, December 2018.

[18] A. P. M. Ramos, L. P. Osco, D. E. G. Furuya et al., "A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices," *Computers and Electronics in Agriculture*, vol. 178, Article ID 105791, 2020.

[19] X. Yin, M. J. Kroff, and J. Goudriann, "Differential effects of day and night temperature on development to flowering in rice," *Annals of Botany*, vol. 77, no. 3, pp. 203–213, 1996.

[20] M. Ray, "Influence of different weather parameters on rice production-A review," *Advances in Life Sciences*, vol. 5, no. 16, pp. 5776–5782, 2016.

[21] W. A. Mattice, "Weather and corn yields," *Monthly Weather Review*, vol. 59, no. 3, pp. 105–112, 1931.

[22] T. E. Epule, J. D. Ford, S. Lwasa, B. Nabaasa, and A. Buyinza, "The determinants of crop yields in Uganda: what is the role of climatic and non-climatic factors?" *Agriculture & Food Security*, vol. 7, no. 1, pp. 1–17, 2018.

[23] C.-C. Chen and C.-C. Chang, "The impact of weather on crop yield distribution in Taiwan: some new evidence from panel data models and implications for crop insurance," *Agricultural Economics*, vol. 33, no. S3, pp. 503–511, 2005.

[24] Y. Vijayalata, V. N. Rama Devi, P. Rohit, and G. S. S. Raj Kiran, "A suggestive model for rice yield prediction and ideal meteorological conditions during crisis," *International Journal of Scientific & Technology Research*, vol. 8, no. 9, 2019.

[25] J. Han, Z. Zhang, J. Cao et al., "Prediction of winter wheat yield based on multi-source data and machine learning in China," *Remote Sensing*, vol. 12, no. 2, p. 236, 2020.

[26] J. Kim, J. Lee, W. Sang, P. Shin, H. Cho, and M. Seo, "Rice yield prediction in South Korea by using random forest," *Korean Journal of Agricultural and Forest Meteorology*, vol. 21, no. 2, pp. 75–84, 2019.

[27] D. S. D. Z. Abeysiriwardena, K. Ohba, and A. Maruyama, "Influence of temperature and relative humidity on grain sterility in rice," *Journal of the National Science Foundation of Sri Lanka*, vol. 30, no. 1-2, pp. 33–41, 2002.

[28] W. M. U. K. Rathnayake, R. P. De Silva, and N. D. K. Dayawansa, "Assessment of the suitability of temperature and relative humidity for rice cultivation in rainfed lowland paddy fields in Kurunegala district," 2016.

[29] W. M. W. Weerakoon, A. Maruyama, and K. Ohba, "Impact of humidity on temperature-induced grain sterility in rice (*Oryza sativa* L)," *Journal of Agronomy and Crop Science*, vol. 194, no. 2, pp. 135–140, 2008.

[30] S. Peng, J. Huang, J. E. Sheehy et al., "Rice yields decline with higher night temperature from global warming," *Proceedings of the National Academy of Sciences*, vol. 101, no. 27, pp. 9971–9975, 2004.

[31] S. Ratnasiri, R. Walisinghe, N. Rohde, and R. Guest, "The effects of climatic variation on rice production in Sri Lanka," *Applied Economics*, vol. 51, no. 43, pp. 4700–4710, 2019.

[32] M. Esham and C. Garforth, "Agricultural adaptation to climate change: insights from a farming community in Sri Lanka," *Mitigation and Adaptation Strategies for Global Change*, vol. 18, no. 5, pp. 535–549, 2013.

[33] C. S. De Silva, E. K. Weatherhead, J. W. Knox, and J. A. Rodriguez-Diaz, "Predicting the impacts of climate change—a case study of paddy irrigation water requirements in Sri Lanka," *Agricultural Water Management*, vol. 93, no. 1-2, pp. 19–29, 2007.

[34] S.-N. N. Seo, R. Mendelsohn, and M. Munasinghe, "Climate change and agriculture in Sri Lanka: a Ricardian valuation," *Environment and Development Economics*, vol. 10, no. 5, pp. 581–596, 2005.

[35] V. Amaratunga, L. Wickramasinghe, A. Perera, J. Jayasinghe, and U. Rathnayake, "Artificial neural network to estimate the paddy yield prediction using climatic data," *Mathematical Problems in Engineering*, vol. 2020, Article ID 8627824, 11 pages, 2020.

[36] L. Wickramasinghe, R. Weliwatta, P. Ekanayake, and J. Jayasinghe, "Modeling the relationship between rice yield and climate variables using statistical and machine learning techniques," *Journal of Mathematics*, vol. 2021, Article ID 6646126, 9 pages, 2021.

[37] R. Goody, J. Anderson, and G. North, "Testing climate models: an approach," *Bulletin of the American Meteorological Society*, vol. 79, no. 11, pp. 2541–2549, 1998.
[38] E. P. Maurer, B. Levi, T. Pruitt, and P. B. Duffy, "Fine-resolution climate projections enhance regional climate change impact studies," *EOS*, vol. 88, no. 47, 2007.
[39] L. Zhao, O. Keith, E. Bou-Zeid et al., "Global multi-model projections of local urban climates," *Nature Climate Change*, vol. 11, pp. 1–6, 2021.