Hindawi
Journal of Mathematics
Volume 2021, Article ID 6646126, 9 pages
https://doi.org/10.1155/2021/6646126

Research Article

Modeling the Relationship between Rice Yield and Climate Variables Using Statistical and Machine Learning Techniques

Lasini Wickramasinghe, Rukmal Weliwatta, Piyal Ekanayake, and Jeevani Jayasinghe

Faculty of Applied Sciences, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka

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

Received 18 November 2020; Revised 20 December 2020; Accepted 20 January 2021; Published 2 February 2021

Academic Editor: Mehdi Ghatee

This paper presents the application of a multiple number of statistical methods and machine learning techniques to model the relationship between rice yield and climate variables of a major region in Sri Lanka, which contributes significantly to the country's paddy harvest. Rainfall, temperature (minimum and maximum), evaporation, average wind speed (morning and evening), and sunshine hours are the climatic factors considered for modeling. Rice harvest and yield data over the last three decades and monthly climatic data were used to develop the prediction model by applying artificial neural networks (ANNs), support vector machine regression (SVMR), multiple linear regression (MLR), Gaussian process regression (GPR), power regression (PR), and robust regression (RR). The performance of each model was assessed in terms of the mean squared error (MSE), correlation coefficient (R), mean absolute percentage error (MAPE), root mean squared error ratio (RSR), BIAS value, and the Nash number, and it was found that the GPR-based model is the most accurate among them. Climate data collected until early 2019 (*Maha* season of year 2018) were used to develop the model, and an independent validation was performed by applying data of the *Yala* season of year 2019. The developed model can be used to forecast the future rice yield with very high accuracy.

1. Introduction

The ability to predict the future crop yield facilitates the responsible authorities to take the most appropriate decisions in order to ensure food security. As the human population continues to increase, which requires the efficient utilization of lands, enhancing the yield would be more important than increasing the farming area. Climate is one of the primary factors beyond human control, which determines the crop yield. In this sense, modeling and prediction of crop yield by considering the climate variables has become an interesting research topic.

Use of statistical techniques as well as machine learning algorithms to identify the relationship between past climate variables and yield data is presented in the literature [1–3]. As rice is a primary source of food for more than half the world's population, numerous research approaches were proposed for predicting the rice yield [4]. Similar research studies were conducted to model the relationship between the climatic factors and the yield of some other crops such as barley [5], corn [6], sugar cane [7], citrus [8], tea [9], coconut [10], sorghum [11], maize, and soybean [12]. A multiple number of climatic factors were considered in such research studies for the application of statistical methods and machine learning techniques.

Regression techniques, support vector machines (SVMs), and artificial neural networks (ANNs) are some of the techniques applied to model the relationship between rice yield and climate variables. ANN was applied with some climate parameters (precipitation, minimum temperature, average temperature, and maximum temperature) and reference crop evapotranspiration and yield over four years to predict rice yield in Maharashtra State, India [13]. This model was validated with an accuracy of 97.5%, a sensitivity of 96.3%, and specificity of 98.1% by developing a multilayer perceptron neural network. A similar research work performed on the data of several paddy grown areas in Sri Lanka proved that ANN model (with MSE < 0.386) can be used

with less computational time to predict the future paddy yield based on future climatic data [14]. An advanced application of ANN integrated with multiple linear regression (MLR) and penalized regression models for prediction of rice yield based on weather parameters at the west coast of India was presented by Das et al. [15]. Its normalized root mean square error varied between 0.98 and 36.7%. Upland rice yield responses in Sahel, West Africa, were modeled to climate factors, by using several techniques, namely, MLR, boosted tree regression, and ANN [16]. As per the results, ANN outperformed the other two techniques and the research findings concluded that rainfall, not temperature, was the main climate driver of the rice yield in Sahel. A hybrid MLR-ANN model produced better accuracy than the conventional models, namely, ANN, MLR, support vector regression (SVR), k-nearest neighbour (KNN), and random forest (RF) [17].

SVM is another commonly used machine learning technique applied to model the relationship between rice yield and climate variables. The applicability of SVM in determining the relative influence of several climate factors on paddy rice yield in Southwest China was investigated, and it was found that SVMs outperformed ANN and MLR [18]. The relationship between climate variables and rice yield was quantified by applying MLR, principal component analysis, and SVM on 36 years of climate and yield data in Southwest Nigeria [19]. It concluded solar radiation as the climate variable of the highest influence on rice yield, which maximized yield during monsoon and postmonsoon periods. Testing eleven combinations of phenology, climate, and geography data to predict the site-based rice yields in South China using MLR and advanced machine learning methods like backpropagation neural network, SVM, and RF is presented in [20]. It was shown that machine learning methods were more precise than MLR, and SVM produced the highest precision in yield prediction. A hybrid SVR technique was applied to predict rice yield based on the climate and agricultural data in Taiwan from 1995 to 2015, resulting in an average RMSE and R^2 of 60 and 0.996 [21].

Statistical techniques such as MLR and Gaussian process regression (GPR) have also been used for prediction of rice yield. GPR has proven to be more accurate than SVM with an $R^2 > 0.75$ and yield error less than 10%, when they were applied to predict winter wheat yield in China based on climate and soil data [22]. However, the authors could not find any research publication on the application of GPR to build a relationship between rice yield and climatic data. Application of MLR to estimate the yield of crops such as sugar cane [7], citrus [8], and tea [9] was presented in the literature. Ji et al. compared the effectiveness of MLR models with ANN models for rice yield predictions in the mountainous Fujian Province of China [23]. Based on the values of R^2 and the RMSE, they justified the superiority of ANN models ($R^2 = 0.67$, RMSE = 891) for accurate yield prediction over MLR models.

As per the literature, ANN, support vector machine regression (SVMR), and MLR were applied to predict rice yield accurately based on the climate variables. In this paper, a research study conducted to model the relationship between rice yield and climate variables of a major province in Sri Lanka, which contributes significantly to the paddy harvest in the country, is presented. In addition to the aforementioned techniques, GPR, power regression (PR), and robust regression (RR), which have not been used or rarely used, were considered in this research. As rice is the staple food in Sri Lanka, identifying a suitable technique for predicting the yield is important in numerous ways. Section 2 presents a description of the data used for this research and a statistical analysis of them. A brief introduction to the techniques used for modeling and the criteria used for evaluating their performance are also demonstrated. In Section 3, the research findings are presented, results are analyzed, and the proposed models are validated. Finally, conclusions are presented in Section 4.

2. Materials and Methods

2.1. Data Collection. Rice harvest, yield, and climatic data of two districts in Sri Lanka, namely, *Kurunegala* and *Puttalam*, over the last three decades were collected from the Department of Census and Statistics and the Department of Meteorology of Sri Lanka. Paddy yield data of the two major agricultural seasons (*Yala* and *Maha*) were considered. *Yala* season spans from May to August, while *Maha* season spans from the September to March of the following year. Rainfall, minimum temperature, maximum temperature, evaporation, average wind speed (morning and evening), and sunshine hours are the climatic factors considered for modeling. These monthly climatic data except rainfall were averaged for each season in both districts, and they were used with the total rainfall of each season. The nonlinear relationship between paddy yield and the climatic parameters was defined as given in the following equation:

$$\text{rice yield} = \phi\left(\text{RF}, T_{\min}, T_{\max}, E, \text{SH}, \text{AWS}_m, \text{AWS}_e\right), \quad (1)$$

where RF is the rainfall in mm, $T_{\min}$ is the minimum temperature in °C, $T_{\max}$ is the maximum temperature in °C, E is the evaporation in mm, SH is the number of sunshine hours, and AWS_m is the average wind speed in the morning in km/h and AWS_e is that in the evening.

2.2. Analysis of Data. It is important to understand the variation of rice yield and the harvest over the three decades so as to focus on the priority of the study [24]. The data produced values of 3.6 t/ha and 3.96 t/ha for the yield in *Kurunegala Yala* and *Maha* seasons, respectively, and those for the *Puttalam Yala* and *Maha* seasons were found to be 3.63 t/ha and 3.78 t/ha in order (Table 1). Net harvested area in *Kurunegala* district is about 65,000 hectares, while it is only about 10,000 hectares in *Puttalam* district. The highest harvest was produced in *Kurunegala* district during the *Maha* season with a mean of 230.5 tons and a standard deviation of 88 tons, followed by the *Yala* season of the same district with a mean of 132.9 tons and a standard deviation of 68 tons (Figure 1). The harvest in the Puttalam district is much lower than that in Kurunegala district during the corresponding agricultural seasons with a mean of 36.5 tons

Table 1: Summary of the statistics of rice yield data.

Statistics	Actual yield (t/ha)			
	Kurunegala/Maha	*Kurunegala/Yala*	*Puttalam/Maha*	*Puttalam/Yala*
Mean	3.96	3.59	3.78	3.63
Median	3.94	3.62	3.77	3.65
Range	1.49	1.04	1.49	0.92
First quartile ($Q1$)	3.61	3.45	3.53	3.42
Third quartile ($Q3$)	4.13	3.76	4.00	3.83

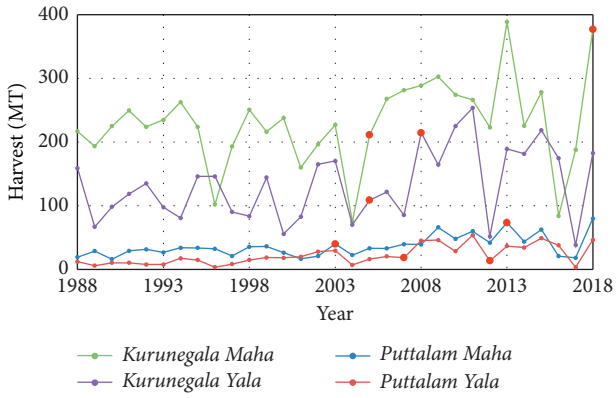

Figure 1: Variation of rice harvest over three decades (1988–2018).

during the *Maha* season and 22.1 tons in the *Yala* season. Therefore, it is evident that the rice yield as well as harvest in the *Maha* season is higher than that in the *Yala* season in each district. A much higher harvest in *Kurunegala* district is obvious due to the huge difference between the net harvested areas. Significant drops in the paddy harvest could be observed in both seasons in the years 2004 and 2012 and in the *Yala* season in year 2017. On the other hand, substantially higher harvests were produced in the *Maha* season in the years 2009 and 2013, *Yala* season in year 2011, and both seasons in year 2018. As the harvest depends on the area under cultivation, which varies with numerous nonclimatic reasons, rice yield data were considered in this research.

The climatic data were also analyzed, and a summary of the statistics is presented in Table 2. As per the analysis of rainfall in both *Kurunegala* and *Puttalam* districts, the average rainfall in *Maha* season was always higher than that in *Yala* season, which must be due to the less precipitation during the southwest monsoon period between May and September and the heavy northeast monsoon that blows in October and November. Further, the average rainfall in *Kurunegala* district is more than that in *Puttalam* district during the corresponding seasons. $T_{\min}$ in corresponding seasons is higher in *Puttalam* district, which lies in three climatic zones (arid, dry, and intermediate), than *Kurunegala* district, which is in the dry zone and intermediate zone of Sri Lanka. However, the highest $T_{\max}$ occurs in *Kurunegala* district during the respective seasons except the minimum of $T_{\max}$ during *Yala* season. The evaporation is greater in *Puttalam* district than in *Kurunegala* district during the comparable *Yala* and *Maha* seasons separately, as evident from the statistics in both districts. This may be due to the influence of more coastal areas in *Puttalam* district associated with strong winds that account for higher evaporation. The figures of the number of sunshine hours suggest that *Kurunegala* district gets more sunshine during both seasons compared to *Puttalam* district. Concurring with basic climatic features, the mean and other associated values in Table 2 indicate that wind speed in the evening is always higher than that in the morning during both seasons of the two districts. Further, stronger winds can be identified during *Yala* season than during *Maha* season in both districts.

2.3. Development of Models. The Levenberg–Marquardt (LM) algorithm, which was developed by combining the gradient descent method and Gauss–Newton method, was used in ANN [25]. A single hidden layer was considered, while climatic data and rice yield data were formed as the data vectors and fed into the ANN model as the input layer and output layer, respectively. In this research study, 70% of the climatic data and the corresponding yield data were used for training, while rest of the data were equally distributed and applied for validation (15%) and testing (15%). Feed-forward network with backpropagation algorithm was used to predict the future yield. Backpropagation is a learning algorithm that enables a network to minimize the error (predicted yield-actual yield) by modifying the weights, which connect neurons (equation (2)). LM algorithm was used in optimization problems due to its faster convergence, and the backpropagation is defined by using the Gauss–Newton method [26]. The behavior of the LM algorithm can be expressed as a relationship between the new weight (x_{k+1}) calculated as a gradient function and the current weight (x_k) determined by using the Newton algorithm.

$$x_{k+1} = x_k - [H + \mu I]^{-1} g, \tag{2}$$

where H is the Hessian calculation approximation, g is the gradient calculation, μ is a constant, and I is an identity matrix.

SVMs are supervised learning methods, which use machine learning theory to improve the accuracy of the prediction models. They can also be used as a regression method (SVMR), keeping all the main features that characterize the maximum margin algorithm [27]. Given a set $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \in (X, Y)$, where X is the set of input vectors and $Y$is the set of output, y is predicted for an unseen $x \in X$. SVMR was used to develop a linear function that can make the best approximation of the dependent response variables. The similarity measures were

Table 2: Summary of the statistics of climatic data.

District/Season	Variable	Mean	Min	Max	Standard deviation	Median
Kurunegala/Maha	Rainfall (mm)	777.4	374	1176	241.0	798
	Minimum temperature (°C)	22.1	20.9	22.7	0.53	22.1
	Maximum temperature (°C)	31.1	30	32.0	0.58	31.2
	Evaporation (mm)	2.61	2.07	3.09	0.26	2.63
	Sunshine (hrs)	6.0	4.4	6.9	0.69	6
	Average morning wind speed (kmph)	1.57	0.92	2.48	0.46	1.55
	Average evening wind speed (kmph)	2.19	1.37	4.4	0.81	1.98
Kurunegala/Yala	Rainfall (mm)	486.7	209.0	882.0	174.9	459.0
	Minimum temperature (°C)	24.1	22.92	24.86	0.53	24.2
	Maximum temperature (°C)	32.1	31.44	33	0.44	32.08
	Evaporation (mm)	3.06	2.68	3.5	0.24	2.99
	Sunshine (hrs)	6.84	6.1	7.6	0.39	6.8
	Average morning wind speed (kmph)	3.28	2.6	4.2	0.51	3.3
	Average evening wind speed (kmph)	3.75	2.4	5.6	1.04	3.5
Puttalam/Maha	Rainfall (mm)	590.8	221	962	229.6	611
	Minimum temperature (°C)	22.9	22.07	23.50	0.39	23.0
	Maximum temperature (°C)	30.5	29.72	31.28	0.43	30.5
	Evaporation (mm)	3.12	2.55	3.8	0.32	3.11
	Sunshine (hrs)	5.18	4.29	5.9	0.49	5.25
	Average morning wind speed (kmph)	4.07	2.4	6.3	1.23	3.6
	Average evening wind speed (kmph)	5.29	3.4	7.7	1.4	4.9
Puttalam/Yala	Rainfall (mm)	331.5	65	675	159.31	316.5
	Minimum temperature (°C)	26.0	25.1	26.88	0.35	26.05
	Maximum temperature (°C)	31.9	31.54	32.76	0.28	31.95
	Evaporation (mm)	4.98	4.68	5.29	0.20	4.95
	Sunshine (hrs)	6.02	4.89	6.97	0.54	6.07
	Average morning wind speed (kmph)	8.11	6.26	10.1	1.2	7.8
	Average evening wind speed (kmph)	10.15	8	12.84	1.52	9.81

done based on dot products as indicated in the following equation:

$$y = \langle w \cdot x \rangle + b, \tag{3}$$

where w and b are regression parameters. During the implementation of SVMR-based models, the input was mapped to a high-dimensional feature space using a kernel function, and then a linear regression model was constructed in the new feature space to achieve two conflicting objectives, namely, minimizing errors and avoiding overfitting. Kernel functions (linear, polynomial, Gaussian, etc.) were incorporated for tuning, and most of the times the Gaussian kernel function outperformed the others.

GPR is a nonparametric machine learning technique, which implements Gaussian processes for regression purposes and works well on nonlinear problems with small sample sizes [28]. It is a modified linear regression model, which explains the response by introducing latent variables. The Gaussian process is a collection of random variables (x, y), whose properties are a finite number of subsets with a joint Gaussian distribution defined as

$$y \sim \mathrm{GP}\left(m(x), k\left(x, x'\right)\right), \tag{4}$$

where the mean function $m(x)$ represents the expectation and the kernel function $k(x, x')$ defines the covariance. It is a collection of random variables $x \in X$ whose properties are any finite number of subsets with a joint Gaussian distribution. The GPR was used in combination with the kernel functions named rational quadratic, squared exponential, Matern 5/2, and exponential. All the aforementioned machine learning techniques-based models (ANN, SVMR, and GPR) were developed in MATLAB environment (version 9.4.0.813654-R2018a).

MLR is a statistical technique, which can be used to model a linear relationship between the input variables (climatic data) and output variable (rice yield) with more than one explanatory variable. MLR also allows determining the overall fit (variance explained) of the model and the relative contribution of each predictor to the total variance explained [29]. It is an extension of ordinary least squares regression and can be expressed as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon, \tag{5}$$

where $n = i$ is the number of observations, β_0 is the y-intercept, which is a constant term, β_n represents the slope coefficients for each input variable, and ε is the model deviation. PR is a nonlinear regression model in which the output (response variable) is modeled in proportion to a power (polynomial) of the inputs (explanatory variables) [30].

$$y = a x_1^b x_2^c \cdots x_i^p, \tag{6}$$

where i is the number of observations and $a, b, c, \ldots, p$ are constants. RR provides an alternative to least squares regression that works with less restrictive assumptions and performs well even when outliers are present in the data [31].

Outliers violate the assumption of normally distributed residuals in least squares regression and tend to distort the least squares coefficients by having more influence than they deserve. Although RR can particularly benefit untrained users, careful consideration should be given as it conducts its own residual analysis and down-weights or completely removes some observations. All the aforementioned statistical models (MLR, PR, and RR) were developed by programming in R software (R 4.0.3).

2.4. Evaluation of Models. The performance of each model was assessed in terms of the mean squared error (MSE), correlation coefficient (R), mean absolute percentage error (MAPE), root mean squared error ratio (RSR) [32], BIAS value [33], and the Nash number.

$$\begin{aligned}
\text{MSE} &= \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N},\\
R &= \frac{\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{\left(\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right)\left(\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right)}},\\
\text{MAPE} &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|x_i - y_i|}{x_i}\right)100\%,\\
\text{RSR} &= \frac{\sqrt{\text{MSE}}}{\sigma_x},\\
\text{BIAS} &= \frac{\sum_{i=1}^{N}(y_i - x_i)}{N},\\
\text{Nash number} &= 1 - \left[\frac{\sum_{i-1}^{N}(x_i - y_i)^2}{\sum_{i-1}^{N}\left(x_i - x_{\text{average}}\right)^2}\right],
\end{aligned} \tag{7}$$

where x is the actual yield, y is the predicted yield, N is the number of data values, and σ_x is the standard deviation of actual yield. In highly accurate models, R and Nash number are closer to 1, while the MSE approaches 0. The negative and positive BIAS values show underestimation and overestimation, respectively, while the values close to zero indicate high model accuracy. Further, the relative standard deviation of the actual yield was calculated to understand the behavior of actual yield data.

$$\text{RSD} = \left(\frac{\text{standard deviation}}{\text{mean}}\right)100\%. \tag{8}$$

3. Results and Discussion

3.1. Results. The climatic and rice yield data were separated into four groups based on the district and the agricultural season and applied on each of the six techniques to identify the relationship among them. The performance of each model was measured in terms of the MSE and correlation coefficient (Figure 2). As per the results, MSE of the models developed for *Maha* season of *Kurunegala* district was less than 0.06 and correlation coefficients were higher than 0.78. However, when the *Yala* season of the same district is considered, five models exhibit correlation coefficients over 0.55, but SVMR-based model demonstrated $R = 0.17$. When *Puttalam* district is considered, MSE of the models corresponding to *Maha* season varied between 0.024 and 0.098. Its correlation coefficients varied between 0.38 and 0.89, while those for *Puttalam Yala* season are higher than 0.71. In terms of MSE, the models corresponding to *Puttalam Yala* season are the best with MSE < 0.033. As per the performance analysis of the models, they were not much accurate, particularly in terms of the correlation coefficient. The less number of source data (climatic data and corresponding rice yield) may be a potential reason for this behavior.

Therefore, having considered the full set of data as a single sample, all the statistical methods and machine learning techniques were applied again to model the relationship between rice yield and climate variables. These results are depicted in Figure 3 and summarized in Table 3. The models resulted in MSE < 0.1, which is comparable to previous results but with higher correlation coefficients. The yield functions obtained by applying statistical methods and derived in terms of the climatic factors are given in equations (9)–(11). Some climatic factors did not appear in the yield function as their impact was minimal compared to rainfall, temperature, and average wind speed.

$$\begin{aligned}
Y_{\text{MLR}} = {} & 7077 - 0.28\text{RF} - 216.5\text{AWS}_m + 199.4\text{AWS}_e\\
& - 139.8T_{\text{min}},
\end{aligned} \tag{9}$$

$$Y_{\text{PR}} = 65078\text{RF}^{-0.039}\text{AWS}_m^{-0.16}\text{AWS}_e^{0.19}T_{\text{min}}^{-0.854}, \tag{10}$$

$$Y_{\text{RR}} = 7372 - 0.284\text{RF} - 176\text{AWS}_m + 171\text{AWS}_e - 153T_{\text{min}}. \tag{11}$$

As per the evaluation performed based on MSE, R, MAPE, Nash number, RSR, and BIAS value, the application of GPR generated the most accurate model among all the statistical methods and machine learning techniques considered in this research. The GPR outperformed the others demonstrating the lowest MSE, MAPE, and RSR. The BIAS value of the GPR model was also very close to zero despite being the second lowest among the six models. Moreover, R and the Nash number of GPR model were closer to 1 compared to the other five techniques.

3.2. Validation. Variation of the rice yield predicted by applying GPR for the period of 2000–2018 was compared with the actual rice yield (Figure 4). The predictions were very close to the actual yield values, with maximum absolute errors of 0.22 t/ha and 0.24 t/ha in *Kurunegala* and *Puttalam* districts, respectively. These very low error values indicate the validity of the GPR model. It was further analyzed by comparing the RSD and MAPE as well. Though the RSD of the actual yield data (9.3%) shows the uncertainty of actual yield in the ensuing years, the GPR model resulted in very lower MAPE (2%), demonstrating the possibility of accurate prediction of yield.

In order to validate the accuracy of the GPR model further, climatic data in year 2019 corresponding to *Yala*

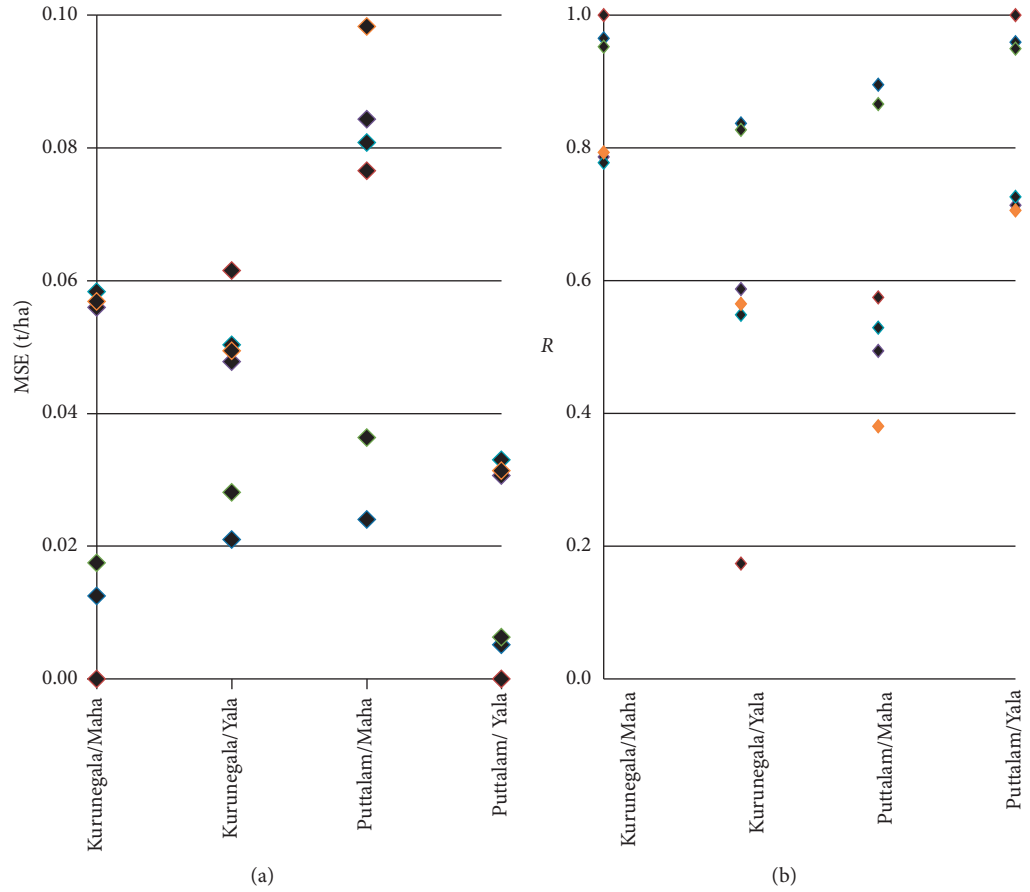

Figure 2: Performance of the models developed separately based on the district and the agricultural season. (a) MSE. (b) R.

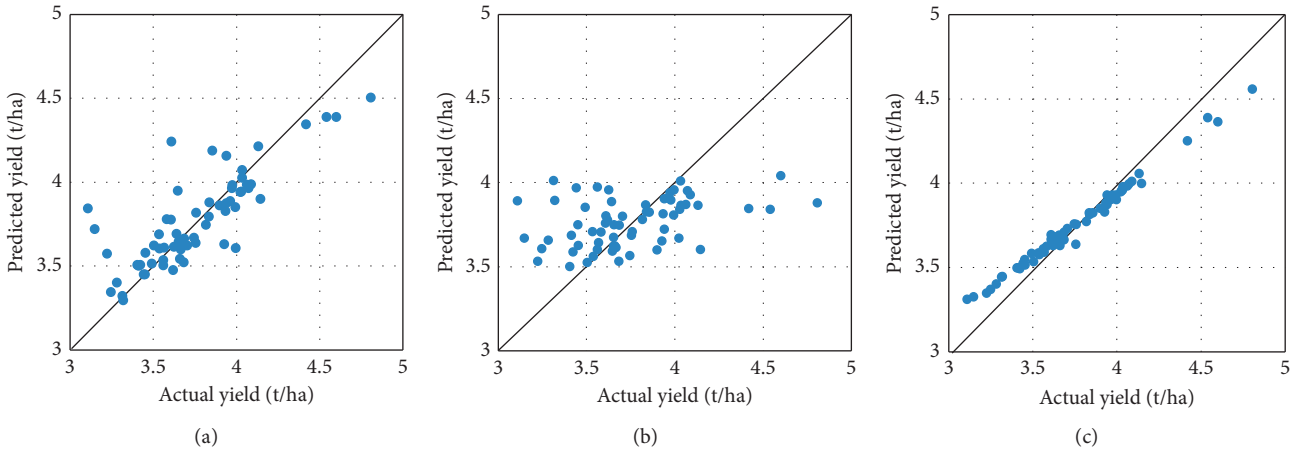

Figure 3: Continued.

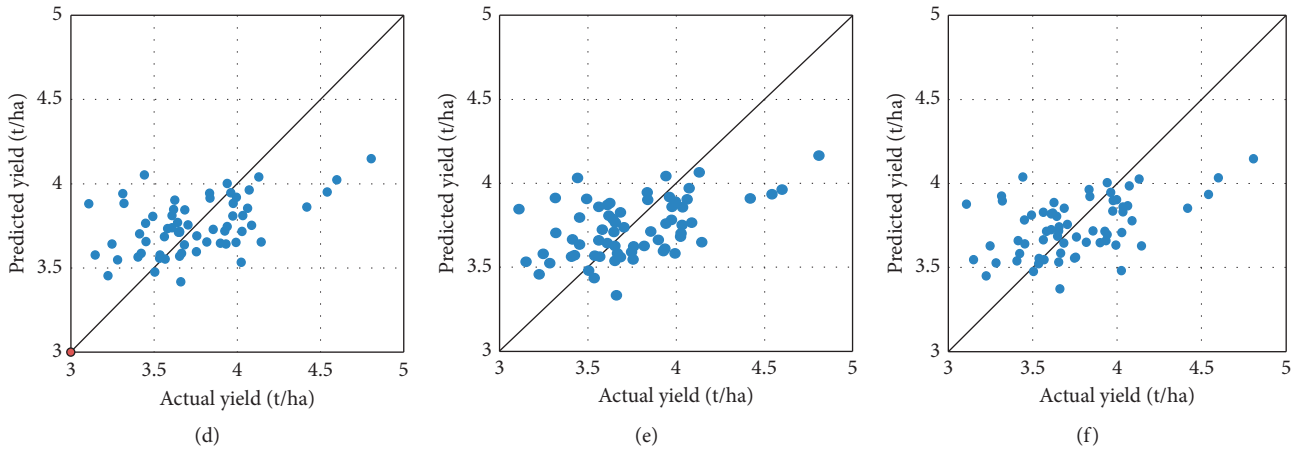

Figure 3: The plot of actual vs. predicted yield after combining all the seasons and districts. (a) ANN. (b) SVMR. (c) GPR. (d) MLR. (e) PR. (f) RR.

Table 3: Performance of the models for the combined dataset.

Parameter	Statistical method/machine learning technique					
	ANN	SVMR	GPR	MLR	PR	RR
MSE	0.04	0.10	0.008	0.09	0.09	0.09
R	0.82	0.24	0.98	0.50	0.51	0.50
MAPE (%)	3.7	6.3	2.0	6.5	6.4	6.4
Nash number	0.67	0.16	0.93	0.25	0.27	0.25
RSR	0.573	0.906	0.256	0.86	0.86	0.86
BIAS	0.023	0.026	0.000001	0.00000006	−0.012	−0.00055

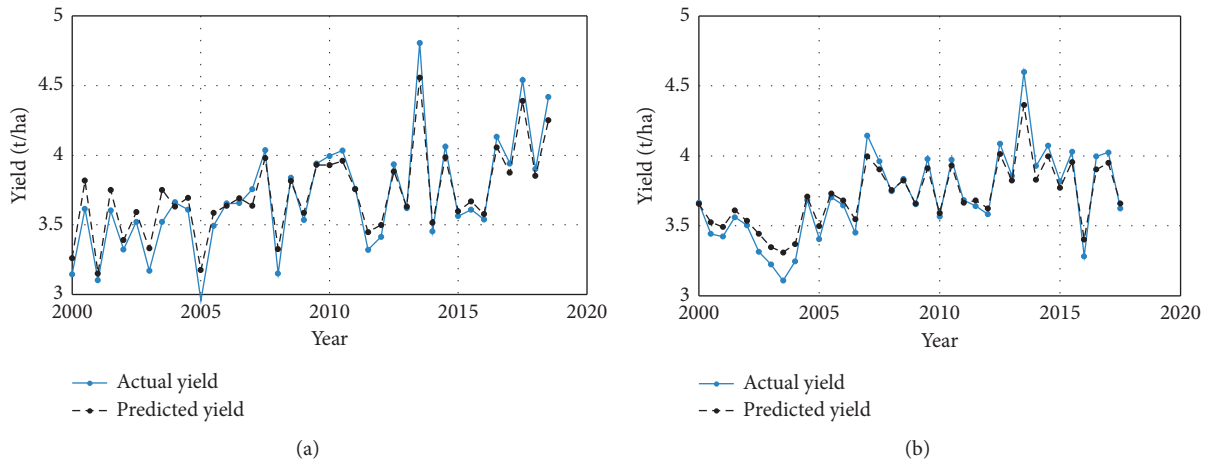

Figure 4: Comparison of actual vs. GPR-based predicted rice yield (2000–2018). (a) *Kurunegala* district. (b) *Puttalam* district.

season of the *Kurunegala* district were applied on the GPR-based model. The predicted rice yield was 3.77 t/ha, while the actual yield was 4.01 t/ha, resulting in an insignificant absolute error of 0.24 t/ha.

4. Conclusions

The relationship between the rice yield and climate variables of two geographically adjacent districts in Sri Lanka was modeled by applying three statistical methods and three machine learning techniques. Based on the significance of weightages and exponents associated with the climatic parameters in the yield functions of the statistical models, the rainfall, temperature, and average wind speed were found to be the most influential climatic factors. The accuracy of the models was evaluated in terms of the MSE, correlation coefficient, MAPE, RSR, BIAS value, and the Nash number. All the machine learning techniques (ANN, SVMR, and GPR) outperformed statistical methods (MLR, PR, and RR) in developing accurate relationship models. The GPR-based

model was the most accurate having MSE, MAPE, RSR, and BIAS values closer to zero while the Nash number and the correlation coefficient approaching 1. Further, an independent validation of the model was conducted by using a recent dataset, which was not used for developing the models. The results demonstrated the capability of the GPR-based model to predict the rice yield by using the known or forecast climatic data.

Data Availability

The data used for the research are available from the corresponding author upon request subject to 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 rice harvest, yield, and climatic data.

References

[1] W. Shi, F. Tao, and Z. Zhang, "A review on statistical models for identifying climate contributions to crop yields," *Journal of Geographical Sciences*, vol. 23, no. 3, pp. 567–576, 2013.

[2] D. B. Lobell and M. B. Burke, "On the use of statistical models to predict crop yield responses to climate change," *Agricultural and Forest Meteorology*, vol. 150, no. 11, pp. 1443–1452, 2010.

[3] A. Crane-Droesch, "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture," *Environmental Research Letters*, vol. 13, no. 11, p. 114003, 2018.

[4] G. S. Khush, "Harnessing science and technology for sustainable rice-based production systems," in *Proceedings of the FAO Rice Conference*, Rome, Italy, February 2004.

[5] L. Parviz, "Assessing accuracy of barley yield forecasting with integration of climate variables and support vector regression," *Annales Universitatis Mariae Curie-Sklodowska, Sectio C–Biologia*, vol. 73, no. 1, pp. 19–30, 2019.

[6] J. W. Bauder and G. W. Randall, "Regression models for predicting corn yields from climatic data and management practices," *Soil Science Society of America Journal*, vol. 46, no. 1, pp. 158–161, 1982.

[7] K. Saithanu, P. Sittisorn, and J. Mekparyup, "Estimation of sugar cane yield in the northeast of Thailand with MLR model," *Burapha Science Journal*, vol. 22, no. 2, pp. 197–201, 2017.

[8] P. Cole and P. McCloud, "Salinity and climatic effects on the yields of citrus," *Australian Journal of Experimental Agriculture*, vol. 25, no. 3, pp. 711–717, 1985.

[9] B. Sitienei, S. Juma, and E. Opere, "On the use of regression models to predict tea crop yield responses to climate change: a case of Nandi east, sub-county of Nandi county, Kenya," *Climate*, vol. 5, no. 3, p. 54, 2017.

[10] B. Das, B. Nair, V. Arunachalam et al., "Comparative evaluation of linear and nonlinear weather-based models for coconut yield prediction in the west coast of India," *International Journal of Biometeorology*, vol. 64, no. 7, pp. 1111–1123, 2020.

[11] S. Sridhara, N. Ramesh, P. Gopakkali et al., "Weather-based neural network, stepwise linear and sparse regression approach for rabi sorghum yield forecasting of Karnataka, India," *Agronomy*, vol. 10, no. 11, p. 1645, 2020.

[12] G. Sakurai, T. Iizumi, and M. Yokozawa, "Varying temporal and spatial effects of climate on maize and soybean affect yield prediction," *Climate Research*, vol. 49, no. 2, pp. 143–154, 2011.

[13] N. Gandhi, O. Petkar, and L. J. Armstrong, "Rice crop yield prediction using artificial neural networks," in *Proccedings of the 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, pp. 105–110, IEEE, Chennai, India, July 2016.

[14] 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.

[15] B. Das, B. Nair, V. K. Reddy, and P. Venkatesh, "Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India," *International Journal of Biometeorology*, vol. 62, no. 10, pp. 1809–1822, 2018.

[16] L. Zhang, S. Traore, J. Ge et al., "Using boosted tree regression and artificial neural networks to forecast upland rice yield under climate change in Sahel," *Computers and Electronics in Agriculture*, vol. 166, p. 105031, 2019.

[17] P. S. M. Gopal and R. Bhargavi, "A novel approach for efficient crop yield prediction," *Computers and Electronics in Agriculture*, vol. 165, p. 104968, 2019.

[18] H. Chen, W. Wu, and H. B. Liu, "Assessing the relative importance of climate variables to rice yield variation using support vector machines," *Theoretical and Applied Climatology*, vol. 126, no. 1-2, pp. 105–111, 2016.

[19] P. G. Oguntunde, G. Lischeid, and O. Dietrich, "Relationship between rice yield and climate variables in southwest Nigeria using multiple linear regression and support vector machine analysis," *International Journal of Biometeorology*, vol. 62, no. 3, pp. 459–469, 2018.

[20] Y. Guo, Y. Fu, F. Hao et al., "Integrated phenology and climate in rice yields prediction using machine learning methods," *Ecological Indicators*, vol. 120, p. 106935, 2021.

[21] T. Z. Jheng, T. H. Li, and C. P. Lee, "Using hybrid support vector regression to predict agricultural output," in *Proceedings of the 2018 27th Wireless and Optical Communication Conference (WOCC)*, pp. 1–3, IEEE, Hualien, Taiwan, April 2018.

[22] 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.

[23] B. Ji, Y. Sun, S. Yang, and J. Wan, "Artificial neural networks for rice yield prediction in mountainous regions," *The Journal of Agricultural Science*, vol. 145, no. 3, p. 249, 2007.

[24] P. Statistics, *Agriculture and Environment Statistics Division*, Department of Census and Statistics, Sri Lanka, 2019.

[25] M. S. Dehaj, M. Z. Mohiabadi, and S. M. S. Hosseini, "Prediction of the outlet flow temperature in a flat plate solar collector using artificial neural network," *Environmental Monitoring and Assessment*, vol. 192, no. 12, pp. 1–15, 2020.

[26] A. Raizada, P. Singru, V. Krishnakumar, and V. Raj, "Development of an experimental model for a magnet

orheological damper using artificial neural networks (Levenberg-Marquardt algorithm)," *Advances in Acoustics and Vibration*, vol. 2016, Article ID 7027259, 6 pages, 2016.

[27] A. Kowalczyk, *Support Vector Machines Succinctly*, Syncfusion Inc., Research Triangle Park, NC, USA, 2017.

[28] C. E. Rasmussen and C. K. Williams, *Gaussian Processes For Machine Learning*, p. 248, MIT press, Cambridge, MA, USA, 2006.

[29] N. R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley & Sons, Hoboken, NJ, USA, 1998.

[30] J. Welc and P. J. R. Esquerdo, *Applied Regression Analysis for Business*, Springer Books, Berlin, Germany, 2018.

[31] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, Hoboken, NJ, USA, 2005.

[32] S. Stajkowski, D. Kumar, P. Samui, H. Bonakdari, and B. Gharabaghi, "Genetic-algorithm-optimized sequential model for water temperature prediction," *Sustainability*, vol. 12, no. 13, p. 5374, 2020.

[33] A. Gholami, H. Bonakdari, I. Ebtehaj, M. Mohammadian, B. Gharabaghi, and S. R. Khodashenas, "Uncertainty analysis of intelligent model of hybrid genetic algorithm and particle swarm optimization with ANFIS to predict threshold bank profile shape based on digital laser approach sensing," *Measurement*, vol. 121, pp. 294–303, 2018.