Journal of Applied Data Sciences
Vol. 2, No. 4, December 2021, pp. 135-142

Soil Infiltration Rate Impact on Water Quality Modeled Using Random Forest Regression

Ajang Sopandi [1,*]

[1] Department of Informatics Engineering, Faculty of Engineering, University of Muhammadiyah Tangerang, Indonesia
[1] ajangsopandi@ft-umt.ac.id*
* corresponding author

(Received: August 6, 2021; Revised: September 10, 2021; Accepted: October 7, 2021; Available online: December 31, 2021)

Abstract

In this paper, Infiltration rate of the soil is investigated by using predictive models of Random forest regression and their performance were compared with Artificial neural network (ANN) and M5P model tree techniques. We utilized 132 field measurements comprising this dataset. 88 models were trained using observations, while the remaining 44 were used to validate it. The cumulative time (Tf), the impurity type (It), the impurity concentration (Ci), and the moisture content (Wc) were utilized as input variables, and the rate of infiltration was employed as the output. To evaluate the efficiency of the two modeling methodologies, correlation coefficients we estimated root mean square error (RMSE), mean absolute error (MAE), relative absolute error (RAE), and root relative square error are all terms that may be used to describe errors (RRSE). The random forest regression approach outperforms the other two models when compared to evolutionary data (ANN and M5P model tree). Using a random forest as a model, regression can properly estimate the infiltration rate within a 25% error range. According to the results of the sensitivity research, cumulative time plays an important influence in determining the soil's penetration rate.

Keywords: Soil Infiltration; Water Quality; Random Forest; Data Mining.

1. Introduction

The term "infiltration" refers to the process through which water percolates through the soil. It is a significant parameter in a variety of applications, including hydrological runoff modeling, irrigation management, and watershed modeling. Infiltration capacity varies geographically and temporally as a result of soil heterogeneity, meteorological variables, clogging processes, and temperature fluctuations, among other things [1]. The soil's capacity to hold moisture varies according to its texture. Sand has a larger pore size than clay, which results in a faster infiltration rate and a decreased capacity for water storage [2]. Additionally, runoff prediction is crucial for hydraulic structure building, as well as for water resource planning and management [3]. The reclamation site's infiltration rates were found to be considerably different from those of undisturbed places in the same region. Agriculture depends on infiltration, which has piqued the curiosity of soil and water scientists [4].

Throughout the past several decades, many prediction algorithms effectively used civil and environmental engineering [5,6]. Numerous linear regression techniques were used to estimate and forecast infiltration rates. The process of modeling neural networks requires the specification of a set of parameters that have been established by the user (number of hidden layers, learning speed, momentum and number of iterations). The number of training rounds affects the reliability of neural network models [7,4]. Another disadvantage of backpropagation neural networks is their lack of training local minima phase. In comparison to neural network techniques, approaches based on support vector and Gaussian process regression outperformed neural network techniques and created global minima. In light of the promising performance of Tree This research investigates the predictive abilities Random forest regression, Tree ANN, SVM, GP, and M5-based regression are all examples of machine learning techniques algorithms. for predicting soil infiltration rates.

2. Literature Review

2.1. Regression on a Random Forest

Random forest (RF) is a classification and regression approach that is composed of a succession of prediction trees, each of which is generated using a randomly chosen random vector that is unrelated to the vector input [8]. A numerical value is given to the prediction tree instead of a class label, as is done with the random forest classifier [9]. When using the random forest regression approach, which was used in this study, the tree was formed by selecting one or more variables at each node at random. For each feature combination, bagging is a strategy that includes randomly selecting N replacement samples from the original training set or a randomly chosen subset of the original training set in order to generate unique trees for each feature combination in the training set [10]. Using the bagging (bootstrap sample) example, the training set contains around 67% of the data from the initial training set, implying that approximately a third of the data is trailing behind each tree planted. The phrase "out-of-bag data" refers to data that has been left over from a previous project (out of bootstrap sampling). Predictor trees need the selection of variable sizes as well as the implementation of a pruning process [11]. Several different techniques for variable selection for tree induction have been presented in the literature, the vast majority of which are directly associated with quality metrics (such as reliability and validity). Variable selection criteria such as the Information Gain Ratio and the Gini Index criterion are the most often employed in tree induction, and they are described in detail below. The Gini Index is used as a selection criterion for variables in the random forest regression technique presented in this article, which is a random forest regression approach. The Gini Index is a measure of the impureness of a variable's output value.

By including a range of factors, the random forest regression design enables trees to grow to their maximum depth using only fresh training data. At the time, the tree is not being pruned.This is one of the most significant benefits of random forest regression over other tree modeling approaches, such as the M5 tree model, in terms of accuracy. According to research, the pruning approach used, rather than the size of the variable selection, has an influence on the performance of tree-based algorithms. According to Mayer et al., [12], the generalization error always converges as the tree population grows, even in the absence of tree trimming, and due to the Strong Law of Large Numbers, overfitting is not a concern.

Several user-defined parameters must be specified for random forest regression to be effective: the number of variables utilized to build trees at each node (m) and the number of trees to be planted (k). At each node, the ideal split is determined using just the supplied variable. Therefore the random forest regression process is composed of k trees, where k is the number of trees to be planted, which may be any value specified by the user and is represented by the symbol k. The random forest regression process is composed of k trees in this manner. As a result of the quantitative nature of the output of forest random-based regression, the average generalization error for each numerical predictor may be calculated. We generated random forest predictors by averaging the generalization error over k trees.

2.2. Model tree M5P

The M5 tree model may be thought of as the binary decision tree is built on top of a linear regression function as its foundation (leaf) that is capable of predicting continuous numeric characteristics [13]. For the purpose of building the tree-based model, the approach of divide and conquer is used. The creation of a model tree is divided into two parts that must be completed independently [14]. The first step is to construct a decision tree based on the separation criteria. The tree model M5 algorithm rules are essentially based If you want to calculate the predicted decrease in this error as a consequence of evaluating each attribute at a node, utilize the standard deviation of the class values reaching that node as a proxy for the fault at that node [14, 15]. In the process of splitting data, the data in the child nodes has a smaller standard deviation than the data in the parent node, resulting in the data in the child nodes being

more pure. M5 selects the split that will result in the greatest predicted error reduction after taking into account all feasible splits [16]. In many cases, this division leads to the formation of a gigantic tree-like structure, which is characterized by overfitting. In order to remedy the issue of overfitting, the tree must be retrimmed, for example, by removing the subtree and replacing it with leafy growth. The second phase of the tree model creation procedure entails trimming the tree and replacing it with a linear regression function [17] in order to prevent the tree from becoming overgrown. Tree model generation divides the parameter space into regions in order to build linear regression models for each sector of the parameter space. This allows for the production of a linear regression model for each sector of the parameter space (subspaces).

2.3. Models of artificial neural networks

Neural networks are composed of components that are easily parallelizable. These components are inspired by the neurological system of animals [18]. As is the case in nature, the primary determinant of a network's functioning is the connections between its components. By altering the connection weight values between pieces, neural networks may be taught to execute certain tasks [19]. In general, neural networks are altered, or trained, to produce certain outputs in response to specific inputs. The network is changed in accordance with the output-to-target ratio until the total of the squared discrepancies between the target and output values equals zero. Typically, the network is trained using numerous such target input/output pairs [20]. Following the presentation of each individual input vector, more training is performed to alter the weights and biases of the network as appropriate. Neural networks are used in a variety of domains of application, including pattern recognition, identification, classification, voice, vision, and control systems, among others. Neural networks are used in a variety of domains of application, including pattern recognition, identification, classification, voice, vision, and control systems.

3. Method

The data set comprises 132 observations, of which 88 are utilized for training and 42 for model testing. The cumulative time (Tf), the kind of impurity (It), the impurity concentration (Ci), and the water content (Wc) are regarded as the input data, whereas the rate of infiltration is regarded as the output.

Table. 1. Data sets used for training and testing have certain characteristics.

Parameters for Input	Train Information				Test Results			
	Min.	Max.	Mean	St. Dev.	Min.	Max.	Mean	St. Dev.
T_f	5.5	180	58.144	55.651	5.3	186	51.031	43.32
I_t	1.5	2	1.477	0.504	1.6	2.5	1.548	0.511
C_i	1.5	15	7.057	5.302	1.5	15.2	6.897	4.890
W_c	4.81	13.68	8.342	3.761	3.87	13.87	9.09	3.625

Coefficients of correlation, correlation coefficients, For the purpose of evaluating performance, the metrics mean square root error, mean absolute error, relative absolute error, and root relative square error were all utilized in combination. Multiple modeling strategies are studied and evaluated on the test data set, which is used to generate the models. Unless the ideal values for the user-defined parameters are provided, the M5P tree model, the artificial neural network, and the random forest techniques will not deliver satisfactory prediction performance. The correlation coefficient (CC), root mean square error (RMSE), mean absolute error (MAE), relative absolute error (RAE), and root relative square error (RRSE) values of a large number of user-defined parameters were calculated and compared

to a test dataset using a variety of machine learning algorithms in order to determine the optimal value for each parameter.

Table. 2. Optimal values for user-defined parameters

Classified Used	Parameters Defined by the User
Model Tree M5P	M = 5
ANN	0.3 learn rate, 0.2 velocity, 11 iterations, 5 hidden layers
Random Forest	K = 2, m = 2, I = 100

4. Conclusions and Discussion

It is necessary to employ performance assessment metrics in order to assess the predictive capabilities of the modeling technique. A variety of statistics were calculated from the data set, including correlation coefficients, root mean squared errors, mean absolute errors, relative absolute errors, and relative square root errors, all of which were returned as values. The plots of actual and anticipated infiltration rates from the training and test datasets illustrate the outcomes of the applied modeling approaches. The three methodologies (tree model M5P, artificial neural network). We examined logistic regression with random forest regression. When the data are compared, it is clear that the random forest regression strategy surpasses all others assessment metrics. The random forest approach's significant increase in prediction accuracy implies that it may be used successfully to forecast the influence of water quality on soil infiltration rates.

Table. 3. Parameters used to evaluate performance using the On the training and testing data sets, M5P model trees, ANNs, and random forest regression were used.

Approaches	Train Data					Test Data				
	CC	MAE	RMSE	RAE	RRSE	CC	MAE	RMSE	RAE	RRSE
M5P Tree model	0.895	5.061	9.342	40.630	51.570	0.892	4.751	5.658	55.201	56.548
ANN	0.960	3.851	5.572	30.847	32.000	0.883	4.265	5.712	49.531	58.000
Random forest	0.990	1.578	3.515	12.453	19.32	0.927	3.185	4.854	36.928	48.491

ISSN 2723-6471

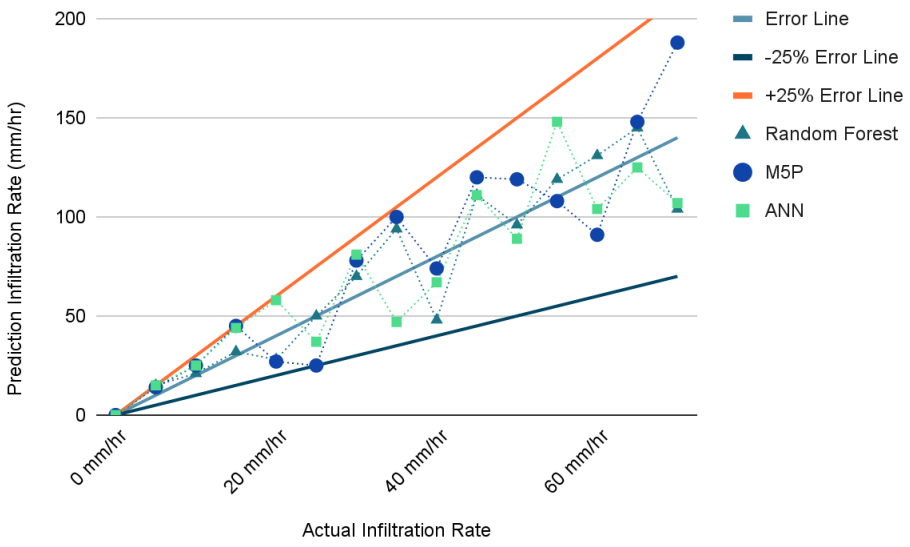

Figure. 1. Actual infiltration rate vs. anticipatedM5P, ANN, and random forest with infiltration rate of training data set

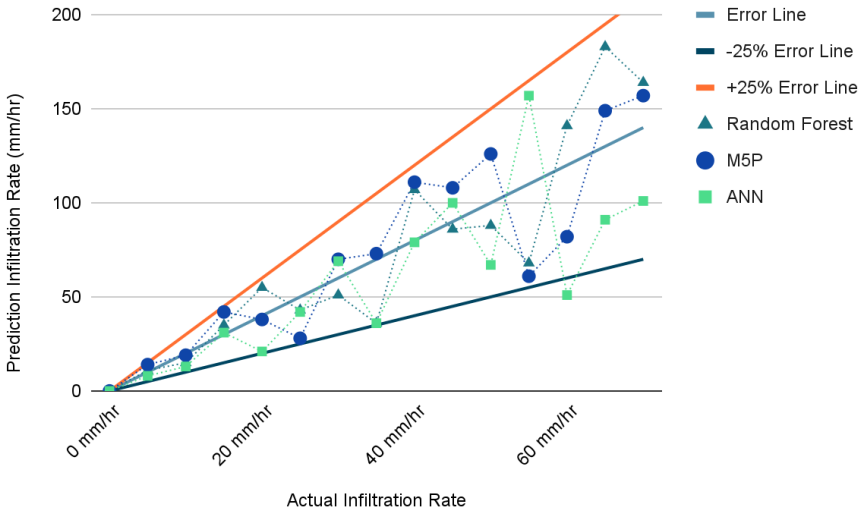

Figure. 2. Actual infiltration rates compared to projections using M5P, ANN, and random forest models.

The graphs in Figures 1 and 2 illustrate the relationship between actual and anticipated infiltration rates using the M5P, AUsing training and testing datasets, we develop NN and random forest tree models. As shown in Figure 1, when the training data set is utilized, the majority, ANN, and random forest are all within a 25% inaccuracy of the perfect agreement line in terms of projected values for each model.

The test dataset, as shown in Figure 2, predicts values that are within a 25% error of the perfect agreement line for M5P, ANN, and Random forest models when employed in conjunction with the test dataset. The ANOVA findings for a single factor indicate that the F value (0.246082) is less than the f-critical value (3.066391) and the P value (0.782224) is more than 0.05, showing that the projected value for the M5P, ANN, and Random tree models is different. Forests are trivial.

The variance in the actual and anticipated infiltration rates as a function of the total number of participants test datasets employing M5P, ANN, and Random forest trees is shown in Figure 3. As demonstrated in this image, the random forest model is capable of accurately predicting the infiltration rate and has a good match to the actual soil infiltration rate. The random forest predicts values in the same way as the genuine values do.

Table. 4. Result of ANOVA Single Factor Tes

Model	F	P-value	F-critical	Distinction in predicted
All Models	0.246	0.783	3.07	Insignificant

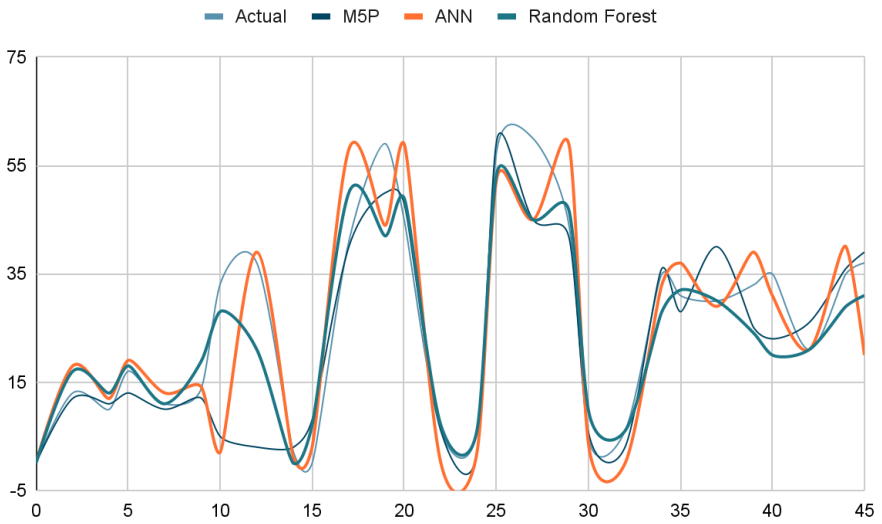

Figure. 3. Variations in the expected infiltration rate compared to the actual infiltration rate of the soil

Table. 5. Random forest regression was used to analyze the sensitivity of the data.

Combination of inputs	Removed input parameter	Regression with a random forest				
		CC	MAE	RMSE	RAE	RRSE
T_f , I_t , C_i , W_C		0.917	3.180	4.851	36.932	48.490
I_t , C_i , W_C	T_f	0.152	11.230	11.817	118.933	118.211
T_f , C_i , W_C	I_t	0.811	4.537	6.791	50.650	67.890
T_f, I_t, W_C	C_i	0.809	5.011	6.970	59.043	69.671
T_f , I_t , C_i	W_C	0.847	4.490	7.672	52.184	76.493

5. Conclusion

The purpose of this research was to determine the predictive capability of computational tools for forecasting the effect of contaminants on the infiltration rate of soils with varied geometries, such as ash and organic fertilizers. The given findings are highly promising and indicate that soil infiltration rates can be analyzed using the M5P, ANN, and Random forest tree models. The Single Factor ANOVA revealed a non-significant difference in predicted values between ANN, tree model M5, and Random Forest. When performance assessment parameters were evaluated, it was discovered that the random forest technique performed better than the M5 tree model and the model considered by ANN for this data set. According to the findings of the sensitivity analysis, cumulative time is the most essential factor in calculating the penetration rate. With increasing cumulative time, the rate of infiltration decreases.

References

[1] H. X. Wu, Y. Zhang, L. Wang, D. Chen, and C. Ma, "Effect of infiltration head on soil water movement of small-diameter tube outflow furrow irrigation under mulch film," World J. Eng., vol. 16, no. 2, pp. 232–237, Jan. 2019, doi: 10.1108/WJE-10-2017-0332.

[2] M. J. Thomas, M. M. Sanjeev, A. P. Sudheer, and J. M.L., "Comparative study of various machine learning algorithms and Denavit–Hartenberg approach for the inverse kinematic solutions in a 3-SS parallel manipulator," Ind. Robot Int. J. Robot. Res. Appl., vol. 47, no. 5, pp. 683–695, Jan. 2020, doi: 10.1108/IR-11-2019-0233.

[3] G. Svensson and C. Padin, "The role of spinoffs and tradeoffs of business-driven sustainable development in the marketplace," J. Bus. Ind. Mark., vol. 36, no. 3, pp. 505–521, Jan. 2021, doi: 10.1108/JBIM-08-2019-0368.

[4] H. Sten Hansen, "Meeting the climate change challenges in river basin planning," Int. J. Clim. Chang. Strateg. Manag., vol. 5, no. 1, pp. 21–37, Jan. 2013, doi: 10.1108/17568691311299345.

[5] S. K. Sarkar, S. Talukdar, A. Rahman, Shahfahad, and S. K. Roy, "Groundwater potentiality mapping using ensemble machine learning algorithms for sustainable groundwater management," Front. Eng. Built Environ., vol. ahead-of-print, no. ahead-of-print, Jan. 2021, doi: 10.1108/FEBE-09-2021-0044.

[6] R. Priyadarshi, A. Panigrahi, S. Routroy, and G. K. Garg, "Demand forecasting at retail stage for selected vegetables: a performance analysis," J. Model. Manag., vol. 14, no. 4, pp. 1042–1063, Jan. 2019, doi: 10.1108/JM2-11-2018-0192.

[7] E. A. Petropoulou, "Indigenous resource management and environmental degradation: southern Greece," Manag. Environ. Qual. An Int. J., vol. 18, no. 2, pp. 152–166, Jan. 2007, doi: 10.1108/14777830710725821.

[8] C. Otchia and S. Asongu, "Industrial growth in sub-Saharan Africa: evidence from machine learning with insights from nightlight satellite images," J. Econ. Stud., vol. 48, no. 8, pp. 1421–1441, Jan. 2021, doi: 10.1108/JES-05-2020-0201.

[9] L. Nogueira de Andrade and M. Garcia Praça Leite, "An analysis of the human activities impact on water quantity in the Jequitinhonha Valley, MG/Brazil," Manag. Environ. Qual. An Int. J., vol. 24, no. 3, pp. 383–393, Jan. 2013, doi: 10.1108/14777831311322677.

[10] V. Nistane and S. Harsha, "Performance evaluation of bearing degradation based on stationary wavelet decomposition and extra trees regression," World J. Eng., vol. 15, no. 5, pp. 646–658, Jan. 2018, doi: 10.1108/WJE-12-2017-0403.

[11] B. B. Mishra, A. Kumar, P. Samui, and T. Roshni, "Buckling of laminated composite skew plate using FEM and machine learning methods," Eng. Comput., vol. 38, no. 1, pp. 501–528, Jan. 2021, doi: 10.1108/EC-08-2019-0346.

[12] M. Mayer, S. C. Bourassa, M. Hoesli, and D. Scognamiglio, "Estimation and updating methods for hedonic valuation," J. Eur. Real Estate Res., vol. 12, no. 1, pp. 134–150, Jan. 2019, doi: 10.1108/JERER-08-2018-0035.

[13] W. K. Loo, "Performing technical analysis to predict Japan REITs' movement through ensemble learning," J. Prop. Invest. Financ., vol. 38, no. 6, pp. 551–562, Jan. 2020, doi: 10.1108/JPIF-01-2020-0007.

[14] L. Liu, Y. Zhao, D. Cheng, and B. Ma, "Soil water and salt distribution characteristics with sand pipe," World J. Eng., vol. 16, no. 1, pp. 44–50, Jan. 2019, doi: 10.1108/WJE-09-2018-0316.

[15] B. Liu, L. Shen, H. You, Y. Dong, J. Li, and Y. Li, "Comparison of algorithms for road surface temperature prediction," Int. J. Crowd Sci., vol. 2, no. 3, pp. 212–224, Jan. 2018, doi: 10.1108/IJCS-09-2018-0021.

[16] S. Lee, H. Ji, J. Kim, and E. Park, "What books will be your bestseller? A machine learning approach with Amazon Kindle," Electron. Libr., vol. 39, no. 1, pp. 137–151, Jan. 2021, doi: 10.1108/EL-08-2020-0234.

[17] S. Kaparthi and D. Bumblauskas, "Designing predictive maintenance systems using decision tree-based machine learning techniques," Int. J. Qual. Reliab. Manag., vol. 37, no. 4, pp. 659–686, Jan. 2020, doi: 10.1108/IJQRM-04-2019-0131.

[18] N. Hernandez, N. Caradot, H. Sonnenberg, P. Rouault, and A. Torres, "Is it possible developing reliable prediction models considering only the pipe's age for decision-making in sewer asset management?," J. Model. Manag., vol. 16, no. 4, pp. 1166–1184, Jan. 2021, doi: 10.1108/JM2-11-2019-0258.

[19] H. Hendricks Franssen, "The impact of climate change on groundwater resources," Int. J. Clim. Chang. Strateg. Manag., vol. 1, no. 3, pp. 241–254, Jan. 2009, doi: 10.1108/17568690910977465.

[20] Y. Gao, K. Chang, X. Xing, J. Liang, N. He, and X. Ma, "Determination of soil water hydraulic parameters from infiltration data," Eng. Comput., vol. ahead-of-print, no. ahead-of-print, Jan. 2021, doi: 10.1108/EC-08-2020-0439.

[21] N. El-Rayes, M. Fang, M. Smith, and S. M. Taylor, "Predicting employee attrition using tree-based models," Int. J. Organ. Anal., vol. 28, no. 6, pp. 1273–1291, Jan. 2020, doi: 10.1108/IJOA-10-2019-1903.

[22] D. A. V. de Paula, R. Artes, F. Ayres, and A. M. A. F. Minardi, "Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques," RAUSP Manag. J., vol. 54, no. 3, pp. 321–336, Jan. 2019, doi: 10.1108/RAUSP-03-2018-0003.

[23] S. Choudhury, D. N. Thatoi, J. Hota, and M. D. Rao, "Predicting crack through a well generalized and optimal tree-based regressor," Int. J. Struct. Integr., vol. 11, no. 6, pp. 783–807, Jan. 2020, doi: 10.1108/IJSI-09-2019-0086.

[24] S. N. Chiemela, F. Noulèkoun, C. J. Chiemela, A. Zenebe, N. Abadi, and E. Birhane, "Conversion of degraded agricultural landscapes to a smallholder agroforestry system and carbon sequestration in drylands," Int. J. Clim. Chang. Strateg. Manag., vol. 10, no. 3, pp. 472–487, Jan. 2018, doi: 10.1108/IJCCSM-08-2015-0116.

[25] A. Adane and W. Bewket, "Effects of quality coffee production on smallholders' adaptation to climate change in Yirgacheffe, Southern Ethiopia," Int. J. Clim. Chang. Strateg. Manag., vol. 13, no. 4/5, pp. 511–528, Jan. 2021, doi: 10.1108/IJCCSM-01-2021-0002.