BAREKENG: Journal of Mathematics and Its Applications
December 2024 Volume 18 Issue 4 Page 2505–2516
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss4pp2505-2516

SOIL MOISTURE PREDICTION MODEL IN PEATLAND USING RANDOM FOREST REGRESSOR

Helda Yunita Taihuttu [1*], Imas Sukaesih Sitanggang[2], Lailan Syaufina[3]

[1,2]Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University
Jl. Meranti, Kampus IPB Dramaga, Bogor 16680, Indonesia
[3]Department of Silviculture, Faculty of Forestry, IPB University
Jl. Meranti, Kampus IPB Dramaga, Bogor 16680, Indonesia

*Corresponding author's e-mail: * **heldayunita@apps.ipb.ac.id***

ABSTRACT

Article History:

Received: 18th, April 2024
Revised: 7th, June 2024
Accepted: 3rd,August 2024
Published:14th, October 2024

Keywords:

Forest and Land Fire;
Prediction Model;
Random Forest Regressor;
Randomized Search CV;
Soil Moisture.

Soil moisture is one of the factors that has recently become the focus of research because it is strongly correlated with forest and land fires, where low soil moisture will increase drought and the incidence of forest and land fires. For this reason, this study aims to create a prediction model for soil moisture as an early prevention of fires in peatlands using the Random Forest Regressor (RFR) algorithm. RFR is used because of its ability to predict values and its resistance to overfitting and outliers. A dataset covering soil moisture, precipitation, temperature, maturity, and peat thickness was collected from August 2019 to December 2023. The data includes soil moisture, precipitation, temperature, maturity, and peat thickness. The data were divided into 80% for modeling and 20% for testing. Model performance was optimized through random search CV, resulting in significant prediction accuracy R-squared: 0.914, MAE: 0.0081, MSE: 0.0007, RMSE: 0 .0271, and MAPE: 0.969. These findings demonstrate the effectiveness of RFR in soil moisture prediction and pave the way for more appropriate and timelier implementation of fire mitigation strategies.

How to cite this article:

H. Y. Taihuttu, I. S. Sitanggang, and L. Syaufina., "SOIL MOISTURE PREDICTION MODEL IN PEATLAND USING RANDOM FOREST REGRESSOR," *BAREKENG: J. Math. & App.*, vol. 18, iss. 4, pp. 2505-2516, December, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Forest and land fires in Indonesia occur almost yearly, impacting public health, environmental damage, economic losses, and disrupting the global climate balance [1]. In particular, fires on peatlands such as those that occurred in Ogan Komiring Ilir Regency, South Sumatra Province, pose more complex challenges because peatlands are formed from dead plant remains and decompose slowly in water-saturated conditions [2]; peatlands are rich in organic materials and susceptible to damage and fire. Peatlands play an essential role in storing carbon in the atmosphere, which can reduce greenhouse gas emissions [3] [4], and are rich in biomass, oxygen production, and biodiversity [4] [5].

The cause of fires on peatlands often occurs due to drought exacerbated by the El Nino event [6]. When precipitation is low and temperatures are high, dryness in peatlands will increase. The thickness and maturity level of peat also influence fires in peatlands [7]. One natural factor that has recently become the focus of research and has been proven to have a significant relationship with the incidence of forest and land fires is soil moisture.

Several previous studies researched to measure the relationship between soil moisture and forest and land fires, soil moisture and forest and land fires such as research [8] which has identified and predicted soil moisture conditions for forest fires in the Iberian Peninsula by using regression analysis with soil moisture, temperature, and forest fire data for each land cover in 2010 and 2014. The results show that increasing temperatures and prolonged drought cause more frequent and more extensive forest fires, the R-squared value = 0.68, and the accuracy of the prediction model is 83.3%. A study [9] Identifies the relationship between soil moisture conditions before the fire season in various land covers based on soil moisture data and historical forest fire events. The result was $r = 0.91$, indicating that more significant fires occurred in the United States on soils with low soil moisture in sagebrush land cover. The Ocean Salinity and Soil Moisture Active Passive (SMAP) data were used to identify the risk of forest fires in Canada using statistical analysis [10]. The study's results show that soil moisture has the potential to be an indicator of increased fires, with a value of $r = 0.93$. A delayed correlation analysis was conducted to determine the correlation between soil moisture levels and fire activity in Australia and California [11]. The results showed a negative correlation between soil moisture and forest fire incidents, where soil moisture decreased as fires increased.

Previous research has shown that soil moisture, influenced by various environmental and climatic factors, is essential in the dynamics of forest and land fires. However, the models used to monitor and predict soil moisture are ineffective because they are limited to fundamental statistical analysis and have yet to be carried out on peatlands.

Predictive approaches, primarily through machine learning technology, can identify complex and non-linear patterns from data to obtain better and more dynamic prediction results. One algorithm in the machine learning method that can be used is Random Forest. The RF algorithm was first introduced in 2001 [12]. RF is used for regression and classification, usually called Random Forest Regressor (RFR) in predictions. RF has several advantages, namely reducing overfitting, handling extensive data, having good tolerance for missing or unbalanced data, being not sensitive to variable scales, and dealing with data that contains noise or outliers [13]. Research [14] uses RFR to predict global soil moisture in the context of the atmosphere and agriculture by combining soil moisture, climate, and vegetation index data. The results show that RFR successfully predicts soil moisture with an RMSE value of 0.05 and a correlation coefficient of 0.9.

Based on the results of previous research, we obtained good results in predicting soil moisture with RFR. For this reason, this research aims to use RFR to predict soil moisture in peatlands using data features such as precipitation, temperature, peat maturity, and peat thickness. Hopefully, this research can contribute to the latest technology in the early prevention of forest and land fires by offering machine learning applications that can function as initial tools for policymakers in formulating land management strategies and improving environmental conservation efforts.

2. RESEARCH METHODS

2.1 Dataset

The data used in this research covers the area of Ogan Komering Ilir Regency, South Sumatra Province, which is one of the districts that frequently experiences forest and land fires. The data and data sources taken in this research are as follows:

1. Soil moisture, precipitation, and temperature data from 1 June 2019 to 31 December 2022 were downloaded from the National Aeronautics and Space Administration (NASA) Power Larc page.
2. Data on the thickness and maturity of peat in Ogan Komering Ilir Regency, South Sumatra Province, from the Center for Agricultural Land Resources Research and Development (BBSDLP) in 2019

2.2 Research Stages

This research consists of several stages: data collection, preprocessing, data partition, model creation, and model evaluation. The research stages can be seen in **Figure 1**.

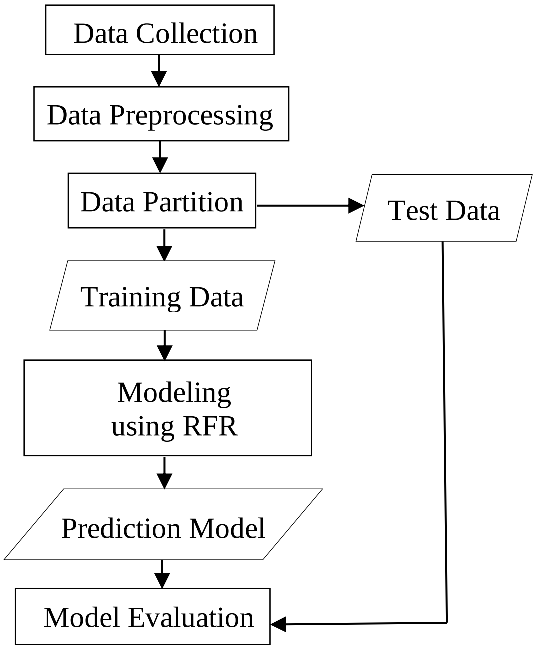


Figure 1. Research Stages

2.2.1 Data Collection

In the initial stage, spatial operations were carried out to obtain coordinate points, which would later be used for data collection at POWER NASA Langley Research Center (LaRC). Map data were processed to obtain peat maturity and thickness data based on coordinate points. This stage was carried out with the help of QuantumGIS. Data were collected from several sources, namely: soil moisture, rainfall, and temperature from June 1, 2019, to December 31, 2023, obtained from POWER LARC NASA (https://power.larc.nasa.gov/data-access-viewer). Meanwhile, peat maturity and peat thickness data were obtained from the peat maps of the Center for Agricultural Land Resources Research and Development (BBSDLP) in 2019.

2.2.2 Data Preprocessing

This stage aims to change the data structure according to the modeling needs. The preprocessing in this research begins with checking and removing outliers. Outlier checking is carried out using box plot visualization, while outlier removal is carried out using the IQR method. The IQR calculation uses **Equation (1) [15]**; when the value is outside the lower and upper limits, the value is considered an outlier and deleted.

Next, the deleted values are considered missing and filled in using the multiple imputation method. This is a powerful and flexible technique for producing more accurate estimates and reducing bias compared to more straightforward methods of filling in missing data, such as the mean or median [16] [17]. Missing data were filled in with the help of the sklearn impute library in Python, using IterativeImputer with Random Forest as an estimator.

$$IQR = Q_3 - Q_1, lower\ limit: Q_1 - 1{,}5\ \times IQR, upper\ limit: Q_3 + 1{,}5\ \times IQR. \tag{1}$$

In the following preprocessing stage, categorical data is transformed into a form that the machine learning model can understand data on peat maturity and peat thickness. In this research, the technique used is ordinal encoding. This technique converts categories into an ordinal scale based on a sequence assigned to an integer value for each category [18].

2.2.3 Dataset Partition

Data division for modeling data is divided into training data and test data. The proportion used is 80% for training data and 20% for test data. Training data is used to carry out modeling using RFR while test data is used to evaluate the prediction model.

2.2.4 Modeling Using RFR

The Random Forest Regressor (RFR) modeling stage was carried out to obtain a model that can predict soil moisture as an early prevention of forest and land fires in Ogan Komering Ilir Regency. At the modeling stage, hyperparameter tuning is carried out using randomized searchCV to find the best parameters for RFR with various values of n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, and Bootstrap. Table 1 shows the clarity of the random forest parameters.

Table 1. Random Forest Parameters

Parameter Name	Information
n_estimators	Number of trees to be built
max_depth	Maximum depth of tree
min_samples_split	Minimum number of samples to separate a node
min_samples_leaf	The minimum number of samples required to become a leaf node
max_features	The number of features considered when finding the best division at each node
Bootstrapping	Sampling technique to determine whether the sample to build each tree was taken with replacement or not

Source: [19]

Randomized search: Search for the best parameters with randomized research using the scikit-learn library in Python. Random Forest (RF) is a machine learning algorithm that belongs to the ensemble learning family. RF is a development of the decision tree concept where random forests combine trees existing in the decision tree into a random forest [13] for better classification and prediction results. The combination technique used is Bootstrap Aggregating or Bagging [12]. The advantage of the bagging method is that it produces individual models that vary from one bootstrap sample to another [20]. This variability reinforces the benefits of the aggregating process, which generally produces more accurate and stable predictions [21]. The final decision result is usually determined by calculating the average of the projections of all the trees in the ensemble [13]. The advantage of random forest is handling missing values, avoiding overfitting, and producing measures of important variables through importance scores [13]. The RF process structure can be seen in Figure 2.

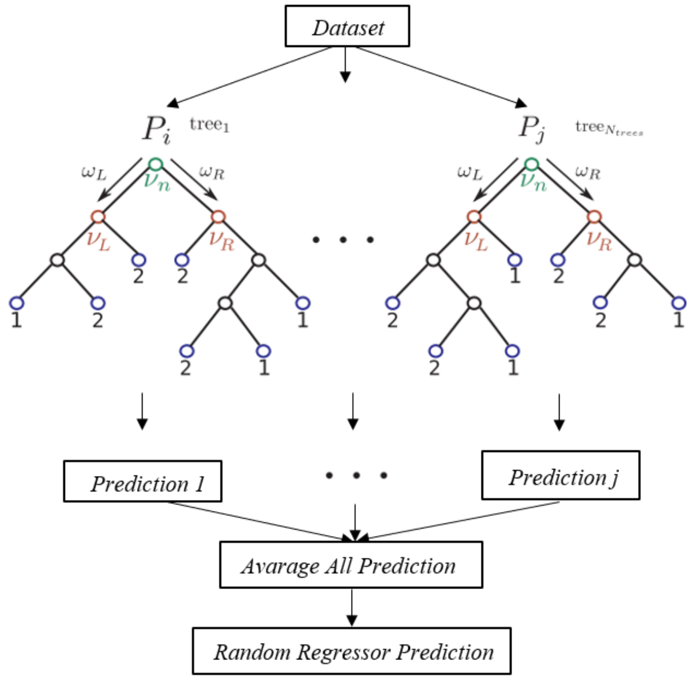


Figure 2. ***Random Forest Regressor*** **Structure (modification of [22])**

Figure 2 shows a schematic representation of the RFR prediction with N_{tree}. The original dataset is used as input to build the RFR model. This dataset contains features that will be used to predict the target values. The formation of decision trees begins by splitting the dataset into several subsets through bootstrap sampling. Each subset is used to build one decision tree. Each tree in the forest is built by splitting the data at each node based on Mean Squared Error (MSE) to find the best split. At each root node (v_L and v_R), statistical measures are used to form homogeneous groups with data partitions (P_i ... P_j) allocated to each tree [22]. w_L and w_R are the proportions of samples decided to go to the left or right branch of the node during the split. These partitions continue to split until reaching terminal nodes (leaf nodes) shown in blue, then each tree provides a prediction of the target value based on the data passing through the nodes within that tree. For example, trees P_i give predictions of 1 and 2 for a particular sample [22]. The predictions from all the trees are combined to provide the final prediction. In the case of regression, the final prediction is the average of all the tree predictions, which can be calculated using Equation (2) [22].

$$\hat{Y}_t = \frac{1}{N_{tree}} \sum_{n=1}^{N_{trees}} \hat{Y}_n \tag{2}$$

where:

$\hat{Y}_t$ = prediction result

N_{trees} = the total number of trees used in the RFR

2.2.5 Model Evaluation

Model evaluation is carried out to evaluate the performance of the prediction model that has been created. Evaluation of the performance of the model used is *the root* determinant coefficient (R^2). R^2 measures how much variation in the target variable can be explained by features. For an optimal prediction model, the R^2 *value* should be approached 1. R^2 can be calculated using Equation (3) [23]. In addition, evaluating model performance also involves the use of error metrics, such as *Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), *and Mean Square Error* (MSE). Values closer to 0 indicate better results, indicating higher prediction errors. Smaller. RMSE, MAE, and MSE can be calculated using Equations (4), (5), and (6) [24]. *Mean Absolute Percentage Error* (MAPE) is also used to assess error by calculating the average absolute value of the percentage difference between the predicted and actual values. MAPE is calculated by Equation (7) [24].

$$R^2 = \frac{\sum_{t=1}^{N}(y_t - \bar{y})^2 - \sum_{t=1}^{N}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{N}(y_t - \bar{y})^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_t - \hat{y}_t| \tag{5}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_t - \hat{y}_t)^2 \tag{6}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100\% \tag{7}$$

where:

$\mathrm{y_t} =$ actual value

$\hat{\mathrm{y}}_\mathrm{t} =$ predicted value

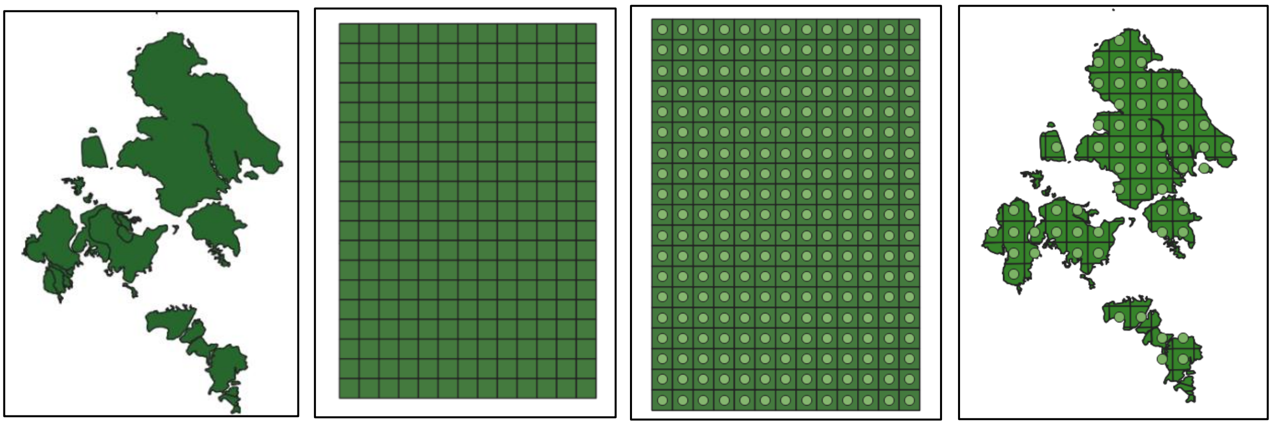
$\bar{\mathrm{y}} =$ The average of the actual values

$n =$ number of predicted data

3. RESULTS AND DISCUSSION

3.1 Data Collection

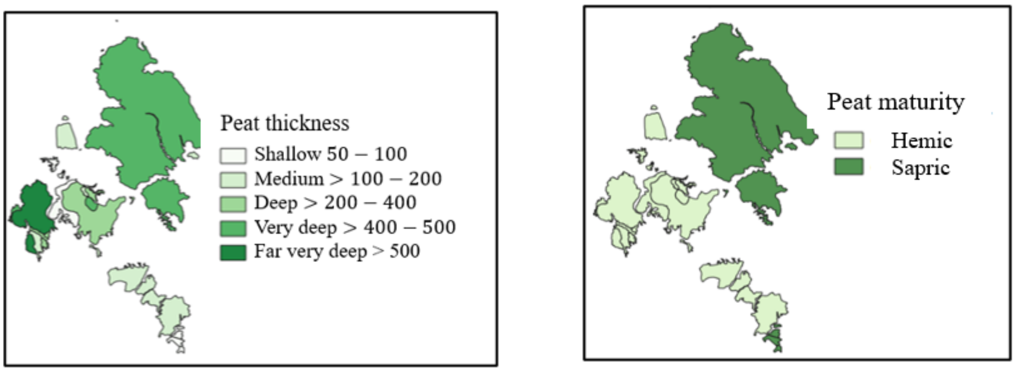
In the initial stage, spatial operations are carried out to obtain coordinate points, which will later be used for data collection. Formation of a grid according to the peatland map of Ogan Komering Ilir Regency using tools in Quantum GIS software. Each grid is represented as a region of size 10 km × 10 km. Next, the centroid of each grid is determined, and then the Intersection operation is carried out to find the centroid within the region. The intersection results show 58 centroid points in the peatland. The stages of grid formation to determine the centroid can be seen in Figure 3.



(a) (b) (c) (d)

Figure 3. (a) Peatland map of Ogan Komering Ilir Regency, (b) Grid 10 ×10, (c) Centroid for each grid, (d) Intersection centroid

Figure 3(d) shows the 58 coordinate points used for data collection (soil moisture, precipitation, and temperature) on NASA's Power Larc. The data were collected daily from August 1, 2019, to December 31, 2023. Meanwhile, peat thickness and maturity data were taken from the 2019 BBSLDP Peat Land Map. Peat thickness and maturity can be seen in Figure 4. Then, the peat thickness and maturity type from each centroid point will be determined with the help of tools such as intersections in QGIS. After that, data on peat thickness and maturity were combined with data on soil moisture, precipitation, and temperature using Microsoft Excel 365. The data were 93,621 rows and five columns (soil moisture, precipitation, temperature, peat maturity, and peat thickness).

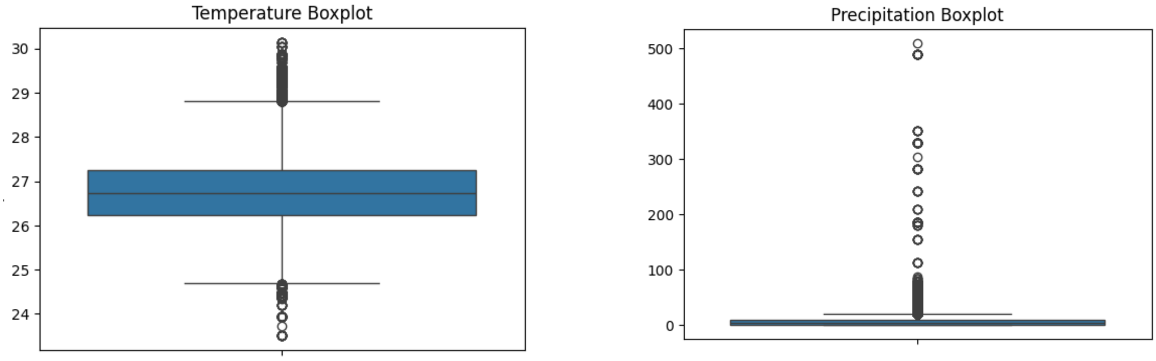


(a) **(b)**

Figure 4. (a) Peatland Map Based on Thickness and (b) Peatland Map Based on Maturity

3.2 Data Preprocessing

Data preprocessing includes checking for outliers using a boxplot (Figure 5), where outliers are removed using the Interquartile Range (IQR) technique and considered missing values. The proportion of missing values is shown in Table 2. Multiple imputation techniques were used with the help of the sklearn to deal with missing values. The impute library in Python uses IterativeImputer with Random Forest as an estimator.



(a) **(b)**

Figure 5. (a) Boxplot of Temperature, (b) Boxplot of Precipitation

Table 2. Presentation of Missing Value

Variables	Missing Value
Precipitation	1.16%
Temperature	12.90%
Soil moisture	3.72%

In the following preprocessing stage, categorical variables such as peat maturity and thickness require transformation using the Ordinal Encoding technique. Peat maturity levels were converted into ordinal values, with hemic = 0 and capric = 1, while thickness: shallow = 0, medium = 1, deep = 2, very deep = 3, and far very deep = 4.

3.3 Data Partition

Data totaling 93,612 rows and five columns that have gone through preprocessing are then divided into 80% training data and 20% test data. Training data is 74,489 and test data is 18,723; data distribution is done randomly.

3.4 Predictive Modeling Using RFR

The soil moisture prediction model using RFR was developed using the sci-kit-learn library in Python. This model involves four feature variables (x): temperature, precipitation, peat maturity, and peat thickness, with soil moisture as the target variable (y). For model parameter optimization, the Randomized SearchCV technique was used in the hyperparameter tuning process, which allows the identification of the best parameter combination through 100 random search iterations, supported by 5-fold cross-validation. Details of the parameters explored in this hyperparameter search are presented in **Table 3**.

Table 3. Parameters in Hyperparameter Search

Parameter	Value
n_Estimator	10 - 200
max_depth	10,20, 30
max_feature	'none', 'sqrt', 'log2'
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
bootstrapping	'True', 'False'

Based on the hyperparameter tuning process using RandomizedSearchCV, the Random Forest Regressor model has been identified with optimal parameters for soil moisture prediction. The best parameters found were as follows: n_Estimator, number of trees in a random forest of 185 with max_depth or maximum depth of each tree of 30, max_features set to 'none', which allows the model to automatically select the best number of features to partition on each node; min_samples_leaf is 1, which means each leaf must have at least one data sample; The min_sample node for splitting a node is set to 2, indicating splitting will only occur if there are at least two samples. The bootstrap method is 'true', meaning samples are selected with replacement when building each tree, thus adding more variation to the model. The prediction model with the best parameter results obtained an R^2 value on test data of 96.1%, showing that the RF model used to predict soil moisture performs very well. A visualization of the comparison of expected results and actual values is presented in **Figure 6**. In contrast, **Figure 7** compares the predicted results and actual values visualized based on coordinate points (latitude and longitude).

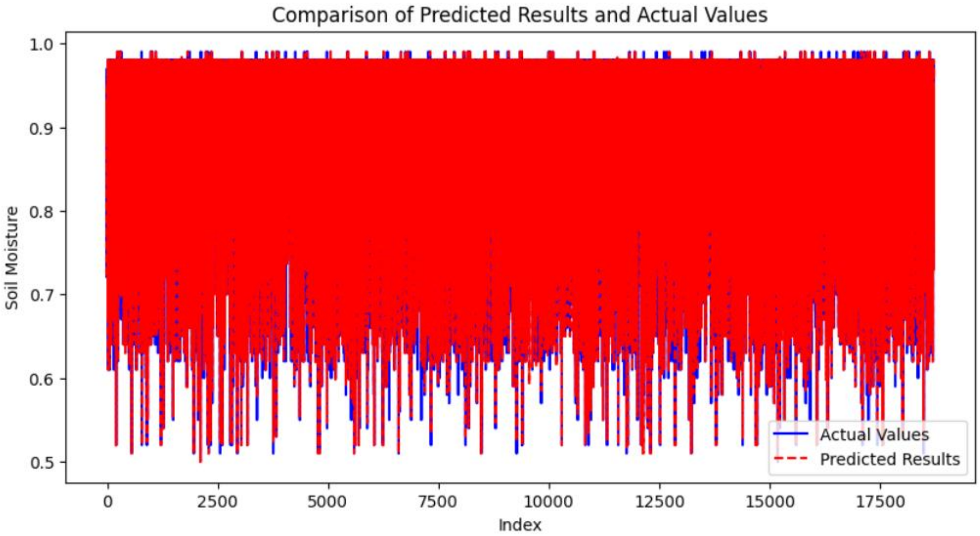
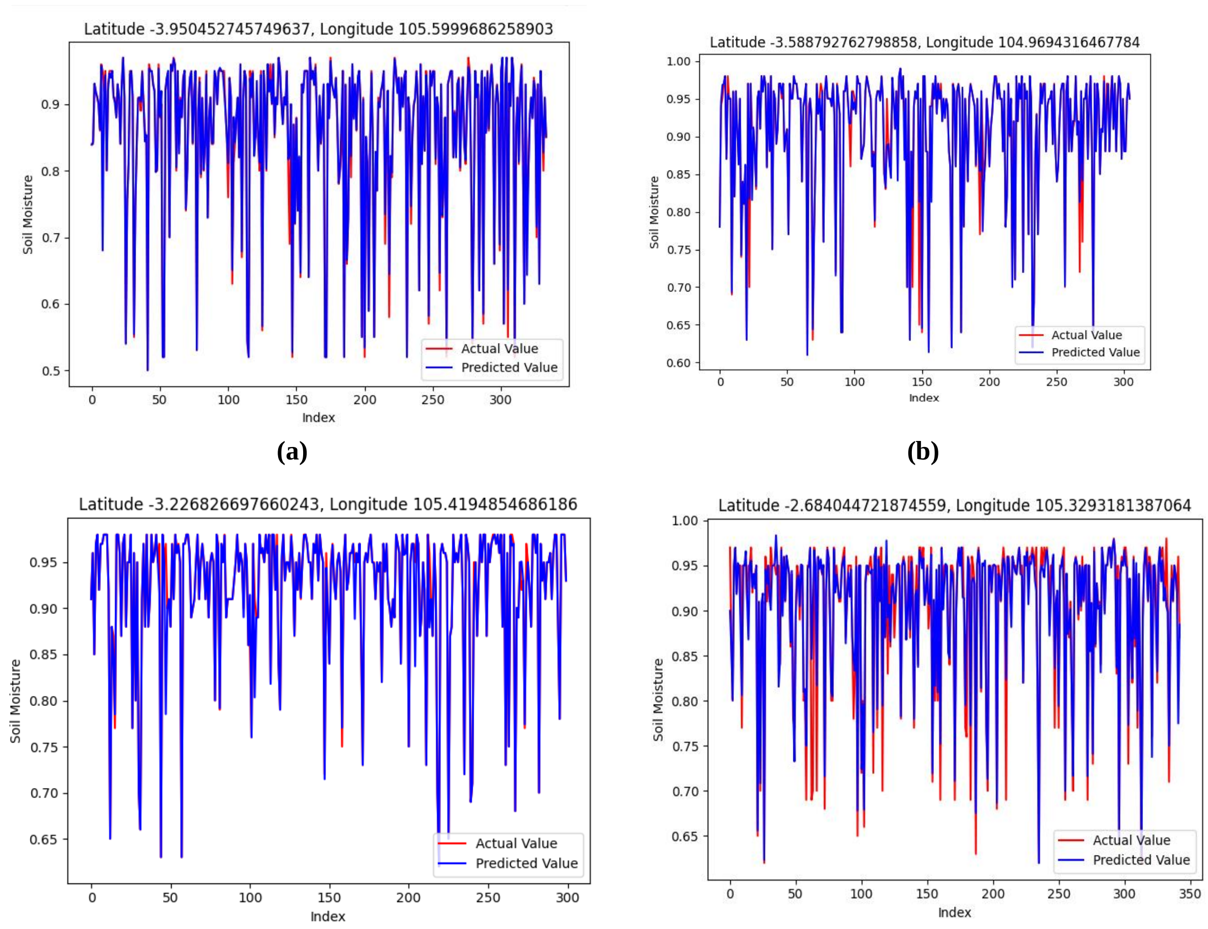


Figure 6. Comparison Plot of Actual and Predicted Values



(a) **(b)**

(c) **(d)**

Figure 7. Comparison Plot of Actual and Predicted Values Based on Coordinate Points (a) Coordinate 1, (b) Coordinate 2, (c) Coordinate 3, and (d) Coordinate 4

Visualization of the comparison of the predicted and the actual value in **Figure 6** and **Figure 7** shows that the prediction results have followed the fluctuations in the exact value. However, some areas of the red line (prediction value) still need to align entirely with the blue line (actual value).

3.5 Evaluation

After the model is built, the model is evaluated using the R^2 evaluation metric, MAPE, MSE, RMSE, and MAE. A comparison of evaluation metric values for training and test data can be seen in **Figure 8** and **Figure 9**.

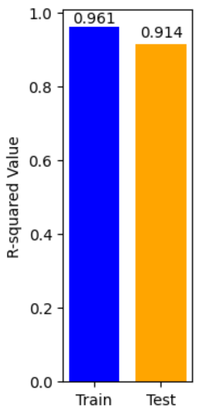


Figure 8. R^2 Value of Prediction Model on Training Data and Test Data

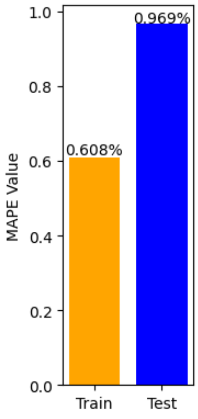


Figure 9. MAPE Value of Prediction Model on Training Data and Test Data

Based on **Figure 8**, the R^2 values for the training and testing data show that the model explains approximately 96.1% of the target variation in the training data, while for the testing data, it is 91.4%. This indicates that the model's ability to explain variability in the testing data is slightly reduced compared to the training data. However, the R^2 values for both datasets remain high. The high R^2 values for both the training and testing data indicate that the soil moisture prediction model using Random Forest Regressor (RFR) is very effective in capturing patterns in the data. The MAPE values for the training and testing data show that the relative error for the training data is 0.608% and for the testing data is 0.969% (**Figure 9**). The low MAPE values for both datasets also indicate that this model has very small prediction errors, making it highly accurate. Evaluation using other error metrics can be seen in **Figure 10**.

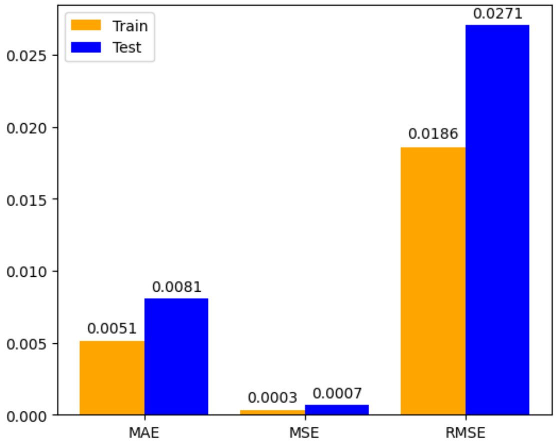


Figure 10. MAE, MSE, and RMSE Values on Training Data and Test Data

Based on **Figure 10**, MAE on training data (0.0051) and test data (0.0081) shows that the model's average absolute error is low. MSE on training data (0.0003) and test data (0.0007) and higher RMSE on test data (0.0186) compared to training data (0.0271) show that the prediction results are very close to the actual value and error variability is minimal.

Overall, the RFR model performs excellently on both training and test data. On test data, the model maintains a high level of accuracy with a reasonable increase in error, indicating that the model generalizes well. The relatively small disparity between training and test data performance suggests that the model does not experience overfitting, and RFR successfully predicts soil moisture well.

4. CONCLUSIONS

This study successfully developed a soil moisture prediction model for peatlands using the Random Forest Regressor (RFR) algorithm, achieving an R-squared value of 0.914 with low prediction errors. This model is effective in predicting soil moisture as a crucial indicator for the early prevention of forest and land fires. It is highly relevant for peatland management and fire mitigation strategies, where soil moisture modeling can be integrated into early warning systems for forest fires. Although the results are promising, the model needs further testing across different types of peatlands and climatic conditions to ensure its generalizability. Future research could focus on integrating additional environmental variables and developing a soil moisture-based early warning system to provide real-time alerts and more effectively aid in forest fire mitigation.

REFERENCES

[1] KLHK, "Identifikasi Areal Bekas Kebakaran Hutan Dan Lahan," Jakarta, 2019. doi: 10.24895/sng.2017.2-0.420.

[2] D. A. P. S. A.F. Falatehan, "Characteristics of Peat Biomass as an Alternative Energy and Its Impact on the Environment," *Solid State Technol.*, vol. 63, no. 5, 2020.

[3] A. Sirin and J. Laine, *Chapter 7 : Peatlands and greenhouse gases*. 2011.

[4] C. Buschmann *et al.*, "Land Use Policy Perspectives on agriculturally used drained peat soils : Comparison of the socioeconomic and ecological business environments of six European regions," *Land use policy*, vol. 90, no. January 2019, p. 104181, 2020, doi: 10.1016/j.landusepol.2019.104181.

[5] L. U. Li *et al.*, "Estimation of Ground Water Level (GWL) for Tropical Peatland Forest Using Machine Learning," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. October, 2022, doi: 10.1109/ACCESS.2022.3225906.

[6] L. Syaufina, I. S. Sitanggang, and L. M. Erman, "Challenges in Satellite-based Research on Forest and Land Fires in Indonesia: Frequent Item Set Approach," *Procedia Environ. Sci.*, vol. 33, pp. 324–331, 2016, doi: 10.1016/j.proenv.2016.03.083.

[7] Fajri, "Model Prediksi Temporal Tinggi Muka Air Sebagai Peringatan Dini Karhutla pada Lahan Gambut Menggunakan LSTM," IPB University, 2022.

[8] D. Chaparro, M. Vall-Llossera, M. Piles, A. Camps, C. Rudiger, and R. Riera-Tatche, "Predicting the Extent of Wildfires Using Remotely Sensed Soil Moisture and Temperature Trends," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 9, no. 6, pp. 2818–2829, 2016, doi: 10.1109/JSTARS.2016.2571838.

[9] D. Jensen, J. T. Reager, B. Zajic, N. Rousseau, M. Rodell, and E. Hinkley, "The sensitivity of US wildfire occurrence to pre-season soil moisture conditions across ecosystems," *Environ. Res. Lett.*, vol. 13, no. 1, 2018, doi: 10.1088/1748-9326/aa9853.

[10] J. T. Ambadan, M. Oja, Z. Gedalof, and A. A. Berg, "Satellite-Observed Soil Moisture As an Indicator of Wildfire Risk," *Remote Sens.*, vol. 12, no. 10, pp. 8–12, 2020, doi: 10.3390/rs12101543.

[11] N. Sazib, J. D. Bolten, and I. E. Mladenova, "Leveraging NASA Soil Moisture Active Passive for Assessing Fire Susceptibility and Potential Impacts over Australia and California," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, no. June, pp. 779–787, 2022, doi: 10.1109/JSTARS.2021.3136756.

[12] L. Breiman, "Random Forests," *Manuf. Netherlands*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[13] V. M. Herrera, T. M. Khoshgoftaar, F. Villanustre, and B. Furht, *Random forest implementation and optimization for Big Data analytics on LexisNexis's high performance computing cluster platform*, vol. 6, no. 1. Springer International Publishing, 2019. doi: 10.1186/s40537-019-0232-1.

[14] L. Zhang *et al.*, "In situ observation-constrained global surface soil moisture using random forest model," *Remote Sens.*, vol. 13, no. 23, pp. 1–25, 2021, doi: 10.3390/rs13234893.

[15] C. Sanjeev, K. Dash, A. Kumar, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decis. Anal. J.*, vol. 6, no. January, p. 100164, 2023, doi: 10.1016/j.dajour.2023.100164.

[16] A. Palanivinayagam and R. Damaševiˇ, "Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods," vol. 14, no. 92, pp. 1–15, 2023.

[17] H. Kang, "The prevention and handling of the missing data," *Korean Soc. Anesthesiol.*, vol. 64, no. 5, pp. 402–406, 2013.

[18] A.-I. Udil, A. Ionescu, and A. Katsifodimos, "Encoding Methods for Categorical Data: A Comparative Analysis for Linear Models, Decision Trees, and Support Vector Machines." 2023. [Online]. Available: http://repository.tudelft.nl/.

[19] E. Luís and P. Da Silva, *"Combining Combining machine learning and deep learning approaches to detect cervical cancer in cytology images."* 2021.

[20] J. K. Afriyie *et al.*, "A Supervised machine learning algorithm for detecting and predicting fraud in credit card transaction," *Decis. Anal. J.*, vol. 100, no. 163, 2023.

[21] G. Louppe, *Understanding Random Forests: From Theory to Practice*, no. October 2014. 2014. doi: 10.13140/2.1.1570.5928.

[22] E. Izquierdo-Verdiguier and R. Zurita-Milla, "An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 88, no. June 2019, 2020, doi: 10.1016/j.jag.2020.102051.

[23] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.

[24] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.