CAUCHY: Jurnal Matematika Murni dan Aplikasi
Volume 10(1) (2025), Pages 442-456
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Geographically Weighted Random Forest Model for Addressing Spatial Heterogeneity of Monthly Rainfall with Small Sample Size

Rismania Hartanti Putri Yulianing Damayanti, Suci Astutik*, Ani Budi Astuti

Department Statistics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Malang, Indonesia

Email: suci_sp@ub.ac.id

ABSTRACT

Rainfall modeling often involves complex spatial patterns that vary across locations. Traditional spatial models such as Geographically Weighted Regression (GWR) assume linear relationships and may fall short in capturing nonlinear interactions among predictors and the small sample size is more challenging to fix the assumptions. To address this limitation, this study applies the Geographically Weighted Random Forest (GWRF) method is a hybrid approach that integrates Random Forest (RF), a non-parametric machine learning algorithm with geographically weighted modeling. GWRF is advantageous as it accommodates both spatial heterogeneity and nonlinear relationships, making it suitable for modeling monthly rainfall, which is inherently spatially varied and influenced by complex factors. This study aims to implement and evaluate the performance of the GWRF model in monthly rainfall prediction across East Java. The model is tested using various numbers of trees to determine the optimal structure, and its performance is assessed using Root Mean Square Error (RMSE), Akaike Information Criterion (AIC), and corrected AIC (AICc). Results indicate that the model tends to overestimate the Out-of-Bag (OOB) Error at all tree variations, with the smallest RMSE (85.68) achieved at 750 trees. Humidity emerges as the most influential variable in predicting monthly rainfall in the region, based on variable importance analysis.

Keywords: Geographically; Heterogeneity Spatial; Rainfall; Random Forest

INTRODUCTION

Geostatistics can be defined as a set of numerical techniques that deal with data involving location attributes. Geostatistics can model spatial trends as well as spatial correlations. Spatial analysis plays an important role in planning, risk assessment, and decision-making in environmental management and conservation [1]. Spatial analysis evaluates a variable geographically which helps in the identification of patterns and trends that may not be apparent from limited point data. There is a well-known spatial analysis method, Geographically Weighted Regression (GWR), introduced by Fotheringham in 2003, used to understand spatial variation in the relationship between dependent and independent variables [2].

Submitted: Pebruary 18, 2025 Reviewed: April 22, 2025 Accepted: May 09, 2025
DOI: https://doi.org/10.18860/cauchy.v10i1.32161

GWR is limited by the assumptions attached to its parameter estimation process, namely the assumptions of linearity and stationarity even with locally varying coefficients. In addition, GWR assumes that the residuals are identical and independent, and does not consider the presence of local multicollinearity that may affect the accuracy of parameter estimation [3]. These limitations can reduce the effectiveness of GWR in capturing the complexity of more complicated spatial data, especially when non-linear relationships or interactions between more complex variables need to be taken into account [4]. Thus, there is a need for an analysis method that is able to capture the complexity of the data.

In recent years, machine learning has developed rapidly and become a modern data analysis method. The advantages of machine learning are flexibility and not limited to linear relationships. Some machine learning methods include Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Gradient Boosting (GB), and Random Forest (RF). Hashimoto et al. (2019) proposed the NASA Earth Exchange Gridded Daily Meteorology (NEX-GDM) RF model to map daily rainfall (among other meteorological variables) with 1 km spatial resolution using satellite, reanalysis, radar, and topographic data for the contiguous United States, from 1979 to 2017 [5]. Previous researchers have compared the performance of several machine learning methods such as Appiah-Badu et al. (2022) conducted research related to rainfall prediction with machine learning algorithms in Ghana [6]. The analysis results show that RF provides better performance than K-NN and Decision Tree. Similarly, research by Nurwatik et al. (2022) concluded that RF is the best model than K-NN and Naïve Bayes for modeling landslide vulnerability in Malang Regency, Indonesia [7]. Other studies have also shown that RF performs well for value prediction [8], [9], [10].

Building on the strengths of RF, researchers have developed Geographically Weighted Random Forest (GWRF) to incorporate spatial heterogeneity into the model. GWRF is based on the concept of a spatially varying coefficient model where the global process is broken down into several local sub-models, similar to the GWR approach [11]. GWRF has the advantage of overcoming multicollinearity problems, so it can process all independent variables without requiring a filtering stage. This model is also able to improve prediction accuracy and provide a more comprehensive analysis of the spatial relationship between independent and dependent variables compared to GWR [3].

Several previous studies have applied GWRF in various cases. Studies on remote sensing and population modeling with 1319 observations show that GWRF can improve prediction accuracy when the spatial scale used is appropriate [12]. In an analysis of the spatial variability of type 2 diabetes mellitus (T2D) prevalence in the United States with 3108 data observations, GWRF models showed that GWRF performed superior to GW-OLS. This model is considered more suitable for spatial analysis, especially in overcoming multicollinearity across different geographic locations [13]. In addition, a study on spatial heterogeneity in traffic accident frequency and its influencing factors in the United States with 18411 observations compared the performance of GW-RF, GWR, and global RF. The results show that GWRF has higher prediction accuracy than global RF, with lower Mean Squared Error (MSE) values, and better overall performance than GWR based on higher R^2 values [14]. Some previous research using big sample size shows that GWRF has a good performance.

Rainfall plays an important role in many aspects, including water resources planning, the agricultural sector, as well as disaster mitigation in East Java [15]. Rainfall variability can affect water availability for irrigation, determine cropping patterns for farmers, and contribute to disaster risks such as floods and droughts. Therefore,

understanding the factors that influence rainfall is essential in order to design more effective adaptation and mitigation strategies in the face of climate change and weather dynamics in the region.

Monthly rainfall is influenced by several key factors, including temperature, humidity and elevation, which can vary spatially [16]. Higher temperatures have the potential to increase evaporation and cloud formation, but under certain conditions can reduce rainfall due to an increase in the atmosphere's capacity to hold water vapor. High relative humidity plays an important role in cloud formation and rainfall, while low humidity levels can inhibit the condensation process and reduce the chance of rain. In addition, elevation affects the distribution of rainfall through the orographic rainfall mechanism, where higher elevation areas tend to receive more rainfall compared to low-lying areas. To understand the pattern of the relationship between monthly rainfall and these factors, this study uses a spatial model approach that is able to capture variations in the pattern of relationships in various locations, so that the results of the analysis can provide more accurate insights in water resources planning and disaster mitigation in East Java. However, the main challenge in this study is the limited rainfall data.

Based on this background, we are interested in studying the GWRF model to overcome spatial heterogeneity in the case of monthly rainfall in East Java with small sample size. The purpose of this study is to apply the GRF model to monthly rainfall, train the model, and explore its performance. In addition, we explore the influence of geographical scale and unique GWRF results, such as the importance of spatial features of the independent variables, to show the impact of the importance of local variables.

METHODS

The research follows several stages: (1) compiling monthly rainfall and climate-related data for East Java, (2) preprocessing and georeferencing the data, (3) building GWRF models with varying numbers of decision trees, (4) evaluating model performance using RMSE, AIC, and AICc, (5) identifying the most influential variables based on variable importance measures.

Data

The data used in this study are secondary data obtained from the East Java Meteorology, Climatology and Geophysics Agency (BMKG) website in November 2023-April 2024 which is the rainy season period. The data used contains three variables, namely daily rainfall (mm), temperature (°C), and air humidity (%). This study uses 11 observation locations as spatial units shown in Figure 1.

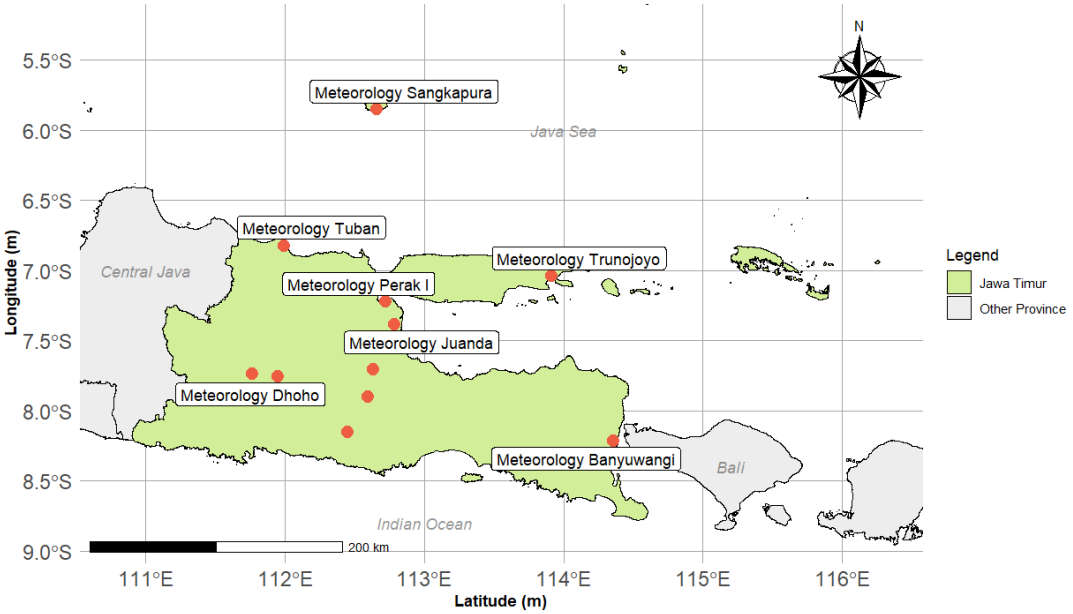

Figure 1. Study Area

The study area in this research is East Java, which is one of the provinces in Indonesia. Astronomically, East Java is located between 111°0' to 114°4' East Longitude and 7°12' to 8°48' South Latitude. East Java Province borders the Java Sea to the north, the Indian Ocean to the south, the Bali Strait to the east, and Central Java Province to the west. This astronomical location gives East Java a tropical climate with weather variations influenced by latitude and altitude from sea level. East Java has 11 weather and climate observation stations. Weather and climate observation stations have a strategic role in providing data that forms the basis for scientific analysis and evidence-based policy making. The resulting long-term data is essential for monitoring climate dynamics, including rainfall patterns, temperature and humidity, to understand trends in environmental change and their impact on ecosystems.

This study uses three variables, including average total rainfall per month, average temperature, and average humidity. Details of the variables used in this study are presented in Table 1.

Table 1. Details of the variables

No.	Variable		Definition	Unit
1	Monthly Rainfall (Y)	Response Variable	The average amount of rainfall each month that falls in an area.	Milimetre (mm)
2	Temperature (^{0}C)	Predictor Variable	A measure of how hot or cold the air is at a given time and place.	Celcius Degree (°C)
3	Humidity (%)	Predictor Variable	The average percentage of water vapor content in the air at a given time and place.	Percent (%)
4	xy	Spatial element	Coordinates of spatial data point	Degree

Exploratory Analysis

To examine the overall relationship between monthly rainfall and its influencing factors, we utilized statistics descriptive, and spatial distribution maps of the factors. Additionally, we computed the Pearson correlation coefficient to quantify the association

between monthly rainfall and the factors. The Pearson correlation formula is shown as follows [17]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

with r as correlation coefficient which has range $[-1,1]$. A negative coefficient indicates the opposite direction of the relationship, while a positive sign indicates a unidirectional relationship. Between two variables x and y with n observations. The strength of the correlation between two variables is categorized as follows Table 2 [17].

Table 2. Categorization of correlation coefficient

Absolute magnitude of the Observed Correlation Coefficient	Interpretation
0.00-0.10	Negligible correlation
0.10-0.39	Weak correlation
0.40-0.69	Moderate correlation
0.70-0.89	Strong correlation
0.90-1.00	Very strong correlation

Random Forest

Random Forest (RF) was first introduced by Leo Breiman in 2001 as an ensemble method that combines the ideas of bagging (bootstrap aggregating) and random subspace selection [18]. It is a collection of decision trees built using the bagging method. In this approach, multiple decision trees are trained on different subsets of the data. Each tree makes a prediction, and the final result is obtained by aggregating the predictions from all trees to improve overall accuracy. The first step in building an RF is performing bootstrap sampling, which involves randomly selecting samples with replacement from the original dataset. From a dataset with N observations, a sample of ≤N is randomly selected, with some data points potentially being selected more than once. Bootstrap sampling is performed for each tree to be built, ensuring that each tree receives a slightly different dataset.

After bootstrap sampling, a decision tree is constructed from the data. However, at each split in the tree, only a subset of the features is considered to determine the best split, a process known as random feature selection, with the Mean Squared Error (MSE) criterion typically used [19]. The splitting at each node continues until a stopping condition is met, such as the maximum tree depth or a minimum number of data points in the leaf node. Each tree provides a prediction, and the final prediction is the aggregation of all tree predictions. In regression, this aggregation is the average of all tree predictions, typically written in:

$$\hat{f}(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad (2)$$

with

$\hat{f}_t(x)$: the prediction in t −tree

T : Number of tree in forest

One of the advantages of RF is its ability to calculate variable importance which measures how much each feature contributes to forming predictions. The RF regression equation is shown in equation below [11].

$$Y_i = ax_i + e, i = 1, \dots, n \quad (3)$$

Y_i define as respon variable for i-th observation and ax_i define as non linear prediction from RF based on all predicor variables x with e is an error in the model. This equation formed using all datset without spatial element.

Geographically Random Forest

Spatial modeling is done because the data shows the relationship between variables is different depending on the location (spatial heterogeneity). The RF algorithm involving spatial elements developed by Georganos et al. (2021) is known as the GRF algorithm. The main concept of GRF is similar to GWR [11]. For each location i, a local RF calculation is performed which only includes a number of observations in the vicinity. This will result in different RF calculations at each sample point which is commonly referred to as the GRF local model. In GRF, equation (3) is expanded so that equation (4) becomes the GRF global model and equation (5) is the GRF local model.

$$Y_i = a_1 X_{i1} + a_2 X_{i2} + e, i = 1,2, \dots, n \quad (4)$$

$$Y_i = a_i(u_i, v_i)x_i + e, i = 1, \dots, n \quad (5)$$

Where $a_i(u_i, v_i)x_i$ is the calibrated RF model prediction at location i where (u_i, v_i) are the coordinate. A sub-model is built for each data location, taking into account only the surrounding observations. The area used by the sub-model is called the neighborhood or kernel, and the maximum distance between the data points and the kernel is called the bandwidth [20]. There are two common types of kernels, namely 'adaptive' and 'fixed'[21]. The use of adaptive kernels is advantageous when the sample density varies across space [22], [23]. The calculation of the gaussian adaptive kernel is obtained by the formulas written in equation (6).

$$w_j(u_i, v_i) = \exp\left(-\left(\frac{d_{ij}}{b_{i(q)}}\right)^2\right) \quad (6)$$

Where $d_{ij} = \sqrt{(u_1 - u_j)^2 + (v_i - v_j)^2}$

d_{ij} is defined as euclidean distance between i and j locations. $b_{i(q)}$ is defined as adaptive bandwidth with neares neighbour is symboled by q.

GRF modeling is based on the heterogeneity of spatial variation so that each location has a different parameter value. The parameter value in GRF is calculated through the impurity value using the variance of the target value (Y) at each node before and after the split. If node S is divided into two subsets, namely left node (S_L) and right node (S_R), then the variance after separation can be calculated by equation (7).

$$Var_{split} = \frac{N_L}{N} Var(S_L) + \frac{N_R}{N} Var(S_R) \quad (7)$$

Where N_L is the number of observations in left node, N_R is the number of observations in rigth node, $var(S_L)$ is the variance of left node after split, and $var(S_R)$ is tahe variance of right node after split. Impurity decrease calculated as variance reduction is written in the following equation.

$$\Delta impurity = var(S) - Var_{split} \quad (8)$$

The greater variance reduction after splitting, the more effective splitting is in improving the homogeneity of smaller nodes. The determination of variable importance is done by summing up the entire impurity reduction ($\Delta impurity$) of each variable across all trees in the forest. Thus, the importance value of a variable X_j is calculated as follows.

$$VI(X_j) = \sum_{t=1}^{T} \sum_{split \in t} \Delta Impurity_{x_j} \quad (9)$$

As an evaluation of the GRF model, there are two approaches that can be used, namely Out-of-Bag (OOB) error and non-OOB error. OOB error is useful for measuring model accuracy by using data that is not selected in the bootstrap process as a test sample without requiring additional validation data. The calculation of OOB error is done with equation (10).

$$OOB\ error = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{y}_i)^2 \quad (10)$$

with Y_i is actual value in i-th location, $\hat{Y}_i$ as predivcted value without i-th observation in training model, and N is number of observation.

Evaluation Performance

Root Mean Square Error (RMSE) is a statistical metric used to measure how well a model's predictions match actual observations. It calculates the square root of the mean squared differences between predicted and actual values [24]. A lower RMSE indicates higher model accuracy. Unlike MSE, RMSE is in the same unit as the target variable, making it more interpretable [25].

Model performance can also be evaluated using the Akaike Information Criterion (AIC) and its corrected version (AICc). AIC helps compare models based on parsimony, while AICc adjusts for small sample sizes. Their formulas are given as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_0 - \hat{z}_0)^2} \quad (11)$$

$$AIC = -2\ln L + 2k \quad (12)$$

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \quad (13)$$

Where L is the model likelihood, n is the number of observations, and k is the number of model parameters.

RESULTS AND DISCUSSION

Exploratory Data Analysis

For our target dataset, the total number of sample data is 11 observations from rain station points in East Java. The size of this dataset is relatively small, so we would like to extrapolate the performance of RF to overcome spatial heterogeneity in small sample sizes. Descriptive statistics showing an overview of the observed data are shown in Table 3.

Table 3. Descriptive Statistics

Variable	Minimum Value	Maximum Value	Mean
Monthly Rainfall (mm)	186.35	498.1	283.048
Temperature (^{0}C)	23.35	29.42	27.60
Humidity (%)	74.64	87.53	80.27

The minimum monthly rainfall in East Java during the study period was 186.35

mm which occurred at Banyuwangi Meteorological Station. While the maximum monthly rainfall of 498.1 mm occurred at Nganjuk Geophysical Station. The minimum temperature occurred at Pasuruan Geophysical Station and the maximum temperature occurred at Silver I Meteorological Station. The lowest air humidity occurs at Malang Geophysical Station and the highest humidity occurs at Pasuruan Geophysical Station. The distribution of average monthly rainfall values, average temperature, and average air humidity in East Java is visually shown in Figure 2.

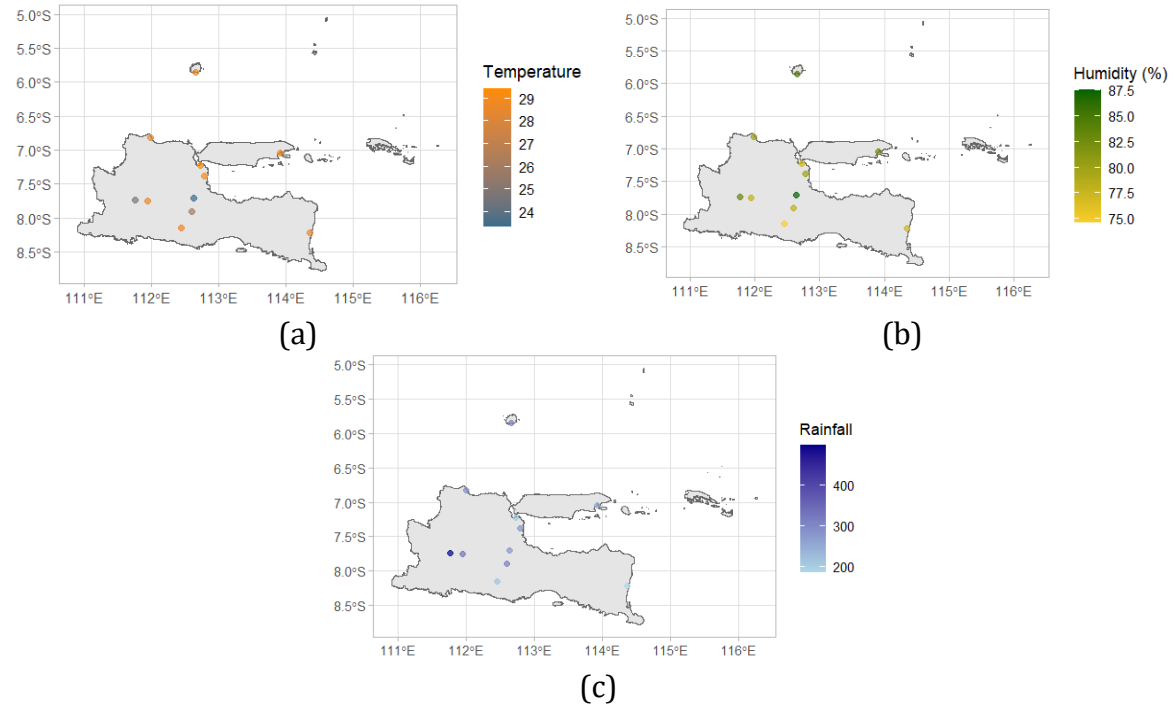

Figure 2. Distribution of (a) temperature, (b) humidity, (c) mothly rainfall in East Java

The spatial distribution of temperature (Figure 2(a)) in the East Java region based on data from several observation points. It can be seen that the distribution of temperature is not homogeneous, but shows variations between locations indicating spatial heterogeneity. Some points in the southwest and northeast parts of the region appear to have relatively higher temperature values than other points, while the central to southern areas tend to show lower temperature values. This distribution pattern may reflect the influence of geographical factors such as altitude, distance from the coast, or land cover that differ between regions [26].

Humidity in the East Java region (Figure (2b)) shows variations between locations, indicating spatial inhomogeneity. Some locations in the northern and central parts show relatively higher humidity levels, while points in the southern and eastern parts tend to have lower humidity. This variation may be influenced by differences in geographical conditions, such as altitude, vegetation and proximity to water bodies. This moisture distribution pattern is important to analyze further as it can affect other atmospheric processes such as cloud formation and rainfall, as well as being one of the important factors in spatial rainfall prediction models.

Figure 3(c) shows the spatial variation of rainfall in East Java, which appears uneven between locations. Points in the west tend to have higher rainfall than other regions. This pattern reflects spatial heterogeneity, which is important to consider in predictive modeling.

Each variable in this study was tested for its relationship with the Pearson

correlation analysis presented in Table 4.

Table 4. Pearson Correlation Analysis

	Temperature	**Humidity**	**Monthly Rainfall**
Temperature	1	-0.5888	-0.5451
Humidity	-0.5888	1	0.4227
Monthly Rainfall	-0.5451	0.4227	1

The results of the Spearman correlation analysis show a significant relationship between temperature, humidity and monthly rainfall. There is a negative correlation of -0.5451 between temperature and monthly rainfall indicating that an increase in temperature tends to reduce rainfall. This is likely because higher temperatures increase the rate of evaporation, thus reducing the amount of rain that falls in an area. Meanwhile, the relationship between humidity and monthly rainfall shows a positive correlation of 0.4227, meaning that the higher the humidity, the greater the likelihood of increased rainfall. High humidity usually indicates more water vapor content in the atmosphere, which has the potential to support cloud formation and rainfall.

The effect of heterogeneity can be identified by the Breusch Pagan test with Breusch Pagan test statistics calculated based on equation (2.1). The results of the test are presented in Table 5.

Table 5. Breusch Pagan Test for Spatial Heterogeneity Test

Statistic	**df**	***P-value***
6.2918	2	0.0430

Based on the test results in Table 5, it is obtained that the p-value is 0.0430 where the value is less than the real level of 0.05. This means that the diversity of rainfall at 11 rainfall stations in East Java is heterogeneous.

Evaluation Performance

The dataset used is 11 examples of observation results from the rainfall station. We determine the response variable is mothly rainfall and predictor variables are temperature and humidity. All data will be useed as training data. Next, we used GRF methods to model the training data using a varying tree numbers in RF algorithm ($k = 10, 50, 100, 500, 750, and\ 100$) as model selection process. The best model of GRF determined based on RMSE, AIC, and AICc. The comparison varying tree numbers (k) in the GRF model is shown in Table 6.

Table 6. Performance Evaluation of Varying Tree Numbers of GRF Model

k	**OOB**			**Not OOB**			**Time(Second)**
	RMSE	**AIC**	**AICc**	**RMSE**	**AIC**	**AICc**	
25	86.32	104.08	107.51	31.10	81.62	85.04	0.9620
50	90.89	105.21	108.64	37.20	85.56	88.98	0.4451
100	96.40	106.51	109.54	34.34	83.80	87.23	0.5055
500	85.98	103.99	107.42	**33.06**	**82.96**	**86.39**	**0.5108**
750	**85.68**	**103.91**	**107.34**	33.87	83.50	86.92	0.6344
1000	86.45	104.11	107.54	33.87	83.50	86.92	1.1043
2000	86.30	104.07	107.50	30.77	81.39	84.82	1.0976
5000	86.24	104.06	107.49	30.43	81.15	84.58	1.7564

The model evaluation results show that varying the number of trees (K) in GWRF from 25 to 5000 does not significantly change the RMSE, AIC, or AICc values. The values of these performance metrics tend to be stable and within a narrow range, which indicates that increasing the complexity of the model through increasing the number of trees does not necessarily improve the prediction accuracy. This may be due to several factors, such as the limited amount of data, the complexity of the relationships between variables that are not too high, or the spatial structure that is not too complex to require a large number of trees. In terms of OOB error, the RMSE values range from 85.68 to 96.40, with the lowest value at k=750. Both AIC and AICc show a similar pattern, with the lowest AIC occurring at k=750 (103.91), and the lowest AICc at k=750 (107.42). For the non-OOB model (Not OOB), the RMSE is lower than the OOB model in all scenarios, with the lowest RMSE (31.10) found at k=25. This happens because the Not OOB model uses the entire dataset for training and evaluation, allowing it to better capture data patterns, while OOB only uses a portion of the data for validation to provide a more objective estimate of model generalization. In terms of computation time, increasing the number of trees raises execution time, with the longest time occurring at k=5000 (1.7564 seconds). Therefore, the optimal number of trees based on accuracy and computational efficiency is k=750.

Feature Importance

The global GRF model builds one model for the entire study area without considering locality differences. The coefficient in the GRF model is the proportion of impurity results on variable importance as presented in Table 7.

Table 7. Global Variable Importance

Variable	Impurity
Temperature	25015.98
Humidity	24257.44

The global feature importance of GRF model places the avaerage of temperature in East Java is more important than humidity in monthly rainfall modeling. Besides globally, GRF is able to identify locally important features shown in Table 8.

The local variable importance results from the GRF model indicate that air humidity generally has a greater contribution than average temperature in predicting monthly rainfall in East Java. This is evident from the higher average importance values for air humidity at most locations. For example, at the Geophysical Station in Nganjuk, the average temperature variable has the highest importance value, while at the Trunojoyo Meteorological Station, its contribution is very low (6242.076). This suggests that temperature may be more relevant in certain areas but less influential in others. Meanwhile, air humidity tends to have more stable importance values, with the highest at the Geophysical Station in Nganjuk (30630.388) and the lowest at the Trunojoyo Meteorological Station (8920.223).

Table 8. Local variable Importance

i	Station	Feature Importance Value	
		Temperature	Humidity
1	Geophysics Malang	26800.056	**27374.475**
2	Climatology Malang	21483.331	**25976.092**
3	Geophysics Nganjuk	29135.592	**30630.388**
4	Geophysics Pasuruan	**21380.199**	21122.283
5	Meteorology Banyuwangi	9483.96	**12994.019**
6	Meteorology Dhoho	**30043.275**	25685.635
7	Meteorology Juanda	15881.001	**17809.18**
8	Meteorology Perak I	15994.245	**18104.771**
9	Meteorology Sangkapura	**16988.595**	12322.308
10	Meteorology Trunojoyo	**6242.076**	8920.223
11	Meteorology Tuban	24950.421	**22031.922**

The importance values indicate the order of importance between the independent variables. Thus, the GRF local model can be written in equation (14).

$$Y_i = a(u_i, v_i)X_{ip} + \epsilon_i, i = 1, \dots, 11, p = 1,2 \quad (14)$$

where,

Y_i : response variable at i location
$\hat{Y}_i$: predicted value
X_{ip} : p-th predictor variable in i location
$a(u_i, v_i)\boldsymbol{X}_i$: predictive function of GRF model
ϵ_i : residual of GRF model

The results of GRF modeling produce a fitted value which is the prediction result at the sample point along with the residuals shown in Table 9.

Table 9. The Results of GRF Model

i	Station	Actual Value (Y_i)	*Fitted Value* ($\widehat{Y}_i$)	Residual (ε_i)
1	Geophysics Malang	216.91	230.86	13.95
2	Climatology Malang	316.07	328.80	12.73
3	Geophysics Nganjuk	498.10	464.57	-33.53
4	Geophysics Pasuruan	283.25	315.66	32.41
5	Meteorology Banyuwangi	186.35	193.03	6.68
6	Meteorology Dhoho	334.89	320.80	-14.10
7	Meteorology Juanda	284.84	276.75	-8.11
8	Meteorology Perak I	206.53	222.12	15.58
9	Meteorology Sangkapura	303.27	309.88	6.61
10	Meteorology Trunojoyo	238.51	252.51	13.99
11	Meteorology Tuban	294.82	291.95	-2.87

The highest residuals occur at Nganjuk Geophysical and Pasuruan Geophysical stations which are areas with the highest monthly rainfall. This indicates that the fitted value of the GRF model has not been able to identify extreme values in the data.

Disccussion

This study applies GWRF modeling to monthly rainfall data in East Java with only 11 observations. The monthly rainfall in East Java exhibits spatial heterogeneity, as indicated by a significant Breusch-Pagan test. The presence of spatial heterogeneity implies that the influence of temperature and humidity varies across different regions, necessitating a localized modeling approach. Previous studies discussed in the introduction section have applied the GWRF model to relatively large sample sizes (>500). This presents a challenge when applying GWRF to small-sample rainfall data [27].

The best GWRF model selection is based on a model selection process with varying numbers of trees. The analysis results show that RMSE, AIC, and AICc decrease as the number of trees increases up to k = 750, after which they increase again at k = 1000. The determination of the optimal number of trees in the RF algorithm cannot be generalized for all cases, as highlighted by Oshiro (2012), who suggested that there may be a threshold where adding more trees leads to diminishing performance while increasing computational costs [28]. The results with different numbers of trees in GRF show that, overall, models using OOB error perform worse than those using Non-OOB error. This indicates that using OOB data for evaluation produces higher error estimates, suggesting that the model struggles to capture patterns in the data. The overestimation of OOB error in the GWRF model may be due to the small sample size, as discussed in study on the overestimation of Random Forest's OOB error [29]. The GWRF model is also capable of identifying variable importance at each location. Based on the variable importance analysis, seven out of the 11 observation sites indicate that humidity is the most important factor in modeling monthly rainfall.

The GWRF model does not require specific distribution assumptions and is able to adjust local weights at each observation location, making it particularly suitable for spatial data with high heterogeneity. For rainfall prediction, such heterogeneity reflects local variations in climate controlling factors that are difficult to capture by global models. Our findings align with previous research that compared spatially weighted approaches to non-spatial methods [30], [31], [32].

CONCLUSIONS

Based on the results of the discussion that has been done, several conclusions related to monthly rainfall modeling are obtained. The analysis shows that there is a spatial heterogeneity effect on monthly rainfall data in East Java which indicates the effect of predictor variables on monthly rainfall is different in each location. There are several key findings, among others, related to the performance of GRF when applied to monthly rainfall data with small sample sizes and how the level of importance of predictor variables at each observation location. The analysis results show that the GRF model overestimates the OOB error, so it can be said that the GRF modeling results are not optimal in small samples. This highlights the need for future studies to incorporate robust validation techniques, such as cross-validation or bootstrapping, especially in spatial datasets with few observations. Furthermore, residuals from the GRF model should be carefully analyzed to detect spatial autocorrelation or non-random patterns that could influence model accuracy. Regarding the importance of predictor variables, the analysis shows that in most of the observed locations, humidity has a higher importance than temperature in predicting monthly rainfall. This finding suggests that rainfall prediction models at a regional level, such as in East Java, should consider locally adaptive methods rather than relying solely on global models that assume homogeneous relationships.

Additionally, improving the interpretability of GWRF models is crucial; in particular, future research should focus on developing or adapting methods to statistically test the significance of variable importance at local levels. These advancements would not only increase the reliability of spatial predictions but also enhance their applicability in decision-making processes, such as water resource management and climate adaptation planning.

This study has limitations related to the exploration of residuals from the GRF that may be suspected to be the cause of overestimation. Thus, future research is expected to be able to handle overestimates in the GRF model by exploring the residuals obtained and can add predictor variables that affect monthly rainfall. In terms of the current GWRF model, it still needs to be refined in its practical application, namely calculating the significance level of variable importance like the global RF method.

REFERENCES

[1] A. Z.- Rutkowska And A. Michalik, "The Use Of Spatial Data Infrastructure In Environmental Management:An Example From The Spatial Planning Practice In Poland," *Environmental Management*, Vol. 58, No. 4, Pp. 619–635, Oct. 2016, Doi: 10.1007/S00267-016-0732-0.

[2] A. S. Fotheringham, C. Brunsdon, And M. Charlton, "Geographically Weighted Regression," *The Sage Handbook Of Spatial Analysis*, Vol. 1, Pp. 243–254, 2009.

[3] M. A. Suprayogi, B. Sartono, And K. A. Notodiputro, "Geographically Weighted Machine Learning Model For Addressing Spatial Heterogeneity Of Public Health Development Index In Java Island," *Barekeng: J. Math. & App.*, Vol. 18, No. 4, Pp. 2577–2588, Oct. 2024, Doi: 10.30598/Barekengvol18iss4pp2577-2588.

[4] A. Sulekan And S. S. S. Jamaludin, "Review On Geographically Weighted Regression (Gwr) Approach In Spatial Analysis," *Malays J Fundam Appl Sci*, Vol. 16, No. 2, Pp. 173–7, 2020.

[5] H. Hashimoto *Et Al.*, "High-Resolution Mapping Of Daily Climate Variables By Aggregating Multiple Spatial Data Sets With The Random Forest Algorithm Over The Conterminous United States," *Intl Journal Of Climatology*, Vol. 39, No. 6, Pp. 2964–2983, May 2019, Doi: 10.1002/Joc.5995.

[6] N. K. A. Appiah-Badu, Y. M. Missah, L. K. Amekudzi, N. Ussiph, T. Frimpong, And E. Ahene, "Rainfall Prediction Using Machine Learning Algorithms For The Various Ecological Zones Of Ghana," *Ieee Access*, Vol. 10, Pp. 5069–5082, 2022, Doi: 10.1109/Access.2021.3139312.

[7] N. Nurwatik, M. H. Ummah, A. B. Cahyono, M. R. Darminto, And J.-H. Hong, "A Comparison Study Of Landslide Susceptibility Spatial Modeling Using Machine Learning," *Ijgi*, Vol. 11, No. 12, P. 602, Dec. 2022, Doi: 10.3390/Ijgi11120602.

[8] H. Mahmoudzadeh, H. R. Matinfar, R. Taghizadeh-Mehrjardi, And R. Kerry, "Spatial Prediction Of Soil Organic Carbon Using Machine Learning Techniques In Western Iran," *Geoderma Regional*, Vol. 21, P. E00260, Jun. 2020, Doi: 10.1016/J.Geodrs.2020.E00260.

[9] R. Taghizadeh-Mehrjardi *Et Al.*, "Multi-Task Convolutional Neural Networks Outperformed Random Forest For Mapping Soil Particle Size Fractions In Central Iran," *Geoderma*, Vol. 376, P. 114552, Oct. 2020, Doi: 10.1016/J.Geoderma.2020.114552.

[10] A. Labade, B. Gupta, R. K. Gupta, And A. Kumar, "Machine Learning-Based Prototype Design For Rainfall Forecasting," In *Machine Intelligence And Data Science*

Applications, A. Ramdane-Cherif, T. P. Singh, R. Tomar, T. Choudhury, And J.-S. Um, Eds., In Algorithms For Intelligent Systems. , Singapore: Springer Nature Singapore, 2023, Pp. 161–172. Doi: 10.1007/978-981-99-1620-7_13.

[11] S. Georganos *Et Al.*, "Geographical Random Forests: A Spatial Extension Of The Random Forest Algorithm To Address Spatial Heterogeneity In Remote Sensing And Population Modelling," *Geocarto International*, Vol. 36, No. 2, Pp. 121–136, Jan. 2021, Doi: 10.1080/10106049.2019.1595177.

[12] S. Georganos And S. Kalogirou, "A Forest Of Forests: A Spatially Weighted And Computationally Efficient Formulation Of Geographical Random Forests," *Ijgi*, Vol. 11, No. 9, P. 471, Aug. 2022, Doi: 10.3390/Ijgi11090471.

[13] S. Quiñones, A. Goyal, And Z. U. Ahmed, "Geographically Weighted Machine Learning Model For Untangling Spatial Heterogeneity Of Type 2 Diabetes Mellitus (T2d) Prevalence In The Usa," *Sci Rep*, Vol. 11, No. 1, P. 6955, Mar. 2021, Doi: 10.1038/S41598-021-85381-5.

[14] S. Wang, K. Gao, L. Zhang, B. Yu, And S. M. Easa, "Geographically Weighted Machine Learning For Modeling Spatial Heterogeneity In Traffic Crash Frequency And Determinants In Us," *Accident Analysis & Prevention*, Vol. 199, P. 107528, May 2024, Doi: 10.1016/J.Aap.2024.107528.

[15] S. Astutik, A. Astuti, R. Damayanti, And A. Syalsabila, "A Hybrid Machine Learning And Kriging Approach For Rainfall Interpolation," *Int. J. Math. Comput. Sci.*, Pp. 271–276, 2025, Doi: 10.69793/Ijmcs/01.2025/Suci.

[16] Y. Andriyana *Et Al.*, "Spatial Durbin Model With Expansion Using Casetti's Approach: A Case Study For Rainfall Prediction In Java Island, Indonesia," *Mathematics*, Vol. 12, No. 15, P. 2304, Jul. 2024, Doi: 10.3390/Math12152304.

[17] P. Schober, C. Boer, And L. A. Schwarte, "Correlation Coefficients: Appropriate Use And Interpretation," *Anesthesia & Analgesia*, Vol. 126, No. 5, Pp. 1763–1768, May 2018, Doi: 10.1213/Ane.0000000000002864.

[18] L. Breiman, "Random Forests," *Machine Learning*, Vol. 45, Pp. 5–32, 2001.

[19] A. Sekulić, M. Kilibarda, G. B. M. Heuvelink, M. Nikolić, And B. Bajat, "Random Forest Spatial Interpolation," *Remote Sensing*, Vol. 12, No. 10, P. 1687, May 2020, Doi: 10.3390/Rs12101687.

[20] C. Brunsdon, A. S. Fotheringham, And M. Charlton, "Spatial Nonstationarity And Autoregressive Models," *Environment And Planning A*, Vol. 30, No. 6, Pp. 957–973, 1998.

[21] M. M. Fischer And A. Getis, Eds., *Handbook Of Applied Spatial Analysis: Software Tools, Methods And Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. Doi: 10.1007/978-3-642-03647-7.

[22] A. Mittal And N. Paragios, "Motion-Based Background Subtraction Using Adaptive Kernel Density Estimation," In *Proceedings Of The 2004 Ieee Computer Society Conference On Computer Vision And Pattern Recognition, 2004. Cvpr 2004.*, Washington, Dc, Usa: Ieee, 2004, Pp. 302–309. Doi: 10.1109/Cvpr.2004.1315179.

[23] G. Jia, A. Tabandeh, And P. Gardoni, "A Density Extrapolation Approach To Estimate Failure Probabilities," *Structural Safety*, Vol. 93, P. 102128, Nov. 2021, Doi: 10.1016/J.Strusafe.2021.102128.

[24] T. O. Hodson, "Root-Mean-Square Error (Rmse) Or Mean Absolute Error (Mae): When To Use Them Or Not," *Geosci. Model Dev.*, Vol. 15, No. 14, Pp. 5481–5487, Jul. 2022, Doi: 10.5194/Gmd-15-5481-2022.

[25] D. Althoff And L. N. Rodrigues, "Goodness-Of-Fit Criteria For Hydrological Models: Model Calibration And Performance Assessment," *Journal Of Hydrology*, Vol. 600, P. 126674, Sep. 2021, Doi: 10.1016/J.Jhydrol.2021.126674.

[26] Y. Feng, C. Gao, X. Tong, S. Chen, Z. Lei, And J. Wang, "Spatial Patterns Of Land Surface Temperature And Their Influencing Factors: A Case Study In Suzhou, China," *Remote Sensing*, Vol. 11, No. 2, P. 182, Jan. 2019, Doi: 10.3390/Rs11020182.

[27] E. Bevacqua, G. Zappa, F. Lehner, And J. Zscheischler, "Precipitation Trends Determine Future Occurrences Of Compound Hot–Dry Events," *Nat. Clim. Chang.*, Vol. 12, No. 4, Pp. 350–355, Apr. 2022, Doi: 10.1038/S41558-022-01309-5.

[28] T. M. Oshiro, P. S. Perez, And J. A. Baranauskas, "How Many Trees In A Random Forest?," In *Machine Learning And Data Mining In Pattern Recognition*, Vol. 7376, P. Perner, Ed., In Lecture Notes In Computer Science, Vol. 7376. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, Pp. 154–168. Doi: 10.1007/978-3-642-31537-4_13.

[29] S. Janitza And R. Hornung, "On The Overestimation Of Random Forest's Out-Of-Bag Error," *Plos One*, Vol. 13, No. 8, P. E0201904, Aug. 2018, Doi: 10.1371/Journal.Pone.0201904.

[30] S. N. Khan, D. Li, And M. Maimaitijiang, "A Geographically Weighted Random Forest Approach To Predict Corn Yield In The Us Corn Belt," *Remote Sensing*, Vol. 14, No. 12, P. 2843, Jun. 2022, Doi: 10.3390/Rs14122843.

[31] Z. Su *Et Al.*, "Modeling The Effects Of Drivers On Pm2.5 In The Yangtze River Delta With Geographically Weighted Random Forest," *Remote Sensing*, Vol. 15, No. 15, P. 3826, Jul. 2023, Doi: 10.3390/Rs15153826.

[32] Y. S. Dewi, S. Hastuti, And M. Fatekurohman, "Analysis Of Stunting In East Java, Indonesia Using Random Forest And Geographically Weighted Random Forest Regression," *Braz. J. Biom.*, Vol. 42, No. 3, Pp. 213–224, Aug. 2024, Doi: 10.28951/Bjb.V42i3.679.