PREDICTING FOREST FIRE HOTSPOTS IN KALIMANTAN USING BEST SUBSET VARIABLE SELECTION, REGULARIZED REGRESSION MODEL, AND BAYESIAN MODEL AVERAGING

***Sri NURDIATI*[1], *I Wayan MANGKU*[1], *Isnayni Feby HAWARI*[1], *Mohamad Khoirun NAJIB*[1]**

DOI: 10.21163/GT_2025.202.10

ABSTRACT

The forest area in Kalimantan continues to decrease due to forest and land fires. One way to prevent this situation in Kalimantan is by predicting the number of hotspots based on climate indicators. Many modeling approaches, such as statistical and machine learning models, can be used. This study uses the best subset selection to build a regression model with regularization and Bayesian Model Averaging (BMA). Several predictors are used to predict the number of hotspots, including precipitation, precipitation anomalies, dry spells, El Nino Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD) indices, and seasonality. The best model is selected based on the performance of the RMSE and R^2 values. The results of the best subset selection obtained are a model consisting of six terms in polynomial form and interactions of precipitation anomalies, dry spells, and IOD index. It can be concluded that there is a significant role of dry spells as a predictor for hotspots due to their presence in almost every term of the equation. The BMA model outperforms the regularization model, with an RMSE value on the test data of 664 hotspots and an R^2 of 88.58%. Although Ridge, LASSO, and Elastic Net perform similarly to the BMA model during the training phase, their reliance on a single model can restrict their ability to generalize to new data. In contrast, BMA offers a more robust and accurate approach by aggregating predictions from multiple models and accounting for uncertainty. This ensemble method enhances BMA's predictive performance on test datasets, making it a valuable tool for accurate forecasting in complex scenarios.

Key-words: *Bayesian model averaging, best subset, machine learning, regression, regularization.*

1. INTRODUCTION

The island of Borneo has a total forest area of 40.8 million hectares (Hardiyanti & Nurmanina, 2020). However, this area is decreasing yearly due to various problems caused by the environment and humans (Margono et al., 2014). Gaveau et al. (2014) estimated that the forest area in Borneo decreased from 558,060 km^2 (75.7%) in 1973 to 389,566 km^2 (52.8%) in 2010, based on satellite imagery. One of the environmental problems that often occurs in Indonesia, especially on the island of Kalimantan, and causes a reduction in forest land is forest fires (Tacconi, 2016). Some examples of the largest forest fires in Indonesia occurred in 1982, 1997-1998, 2015, and 2019 (Najib et al., 2022b; van der Werf et al., 2017). However, there are still many other cases of forest fires that occur every year until now. Forest fires have many negative impacts on the environment and humans, such as material losses, changes in the composition of forest ecosystems, damage to land and forest vegetation, and disruption of human health, especially in communities around the fire location (Borrego et al., 2025; Jolly et al., 2022; Saharjo & Hasanah, 2023). These negative impacts make forest fires potentially threaten the environment and humans, so further prevention and handling measures are needed (Hu et al., 2018).

[1]*School of Data Science, Mathematics and Informatics, IPB University, Bogor, Indonesia, (SN) nurdiati@apps.ipb.ac.id, (IWM) wayanma@apps.ipb.ac.id, (IFH) ifebyhawari@apps.ipb.ac.id, (MKN) mkhoirun@apps.ipb.ac.id*

[*]*Corresponding author: nurdiati@apps.ipb.ac.id*

Climate is one of the factors causing forest fires (Ertugrul et al., 2021). According to Syaufina and Puspitasari (2015), climate conditions such as precipitation, temperature, humidity, and air stability cause the potential for forest fires directly. As an archipelagic country flanked by the Indian Ocean and the Pacific Ocean, climate conditions in Indonesia are greatly influenced by oceanographic conditions, especially in both oceans (Kurniadi et al., 2021; Iskandar et al., 2022). Two types of natural phenomena are global climate variability and can affect oceanographic conditions, i.e., ENSO (El Nino Southern Oscillation) in the Pacific Ocean and IOD (Indian Ocean Dipole) in the Indian Ocean (Rachman et al., 2024; Hidayat et al., 2025). Both phenomena affect climate conditions in Indonesia, especially in terms of precipitation and drought levels that can cause forest fires (Nurdiati et al., 2021).

Climate conditions form specific patterns over a while (Chi et al., 2023). For example, a region with a tropical climate, such as Indonesia, tends to have high precipitation levels during the rainy season and high drought levels during the dry season. This pattern will repeat itself every year following the season period and is usually called a seasonal pattern. The pattern formed by climate conditions over a certain period makes climate conditions predictable for humans. Currently, experts have developed many methods to predict things that will happen in the future based on past data, such as Artificial Intelligence (AI) and machine learning (Latif et al., 2023; Reichstein et al., 2019).

AI and machine learning are extensively applied in various aspects of human life, including predicting future scenarios (Huntingford et al., 2019; Sarker IH, 2021). Experts across fields continuously refine prediction models to improve accuracy and minimize errors (Basha & Rajput, 2019). From education and healthcare to economics, these technologies assist in decision-making by using past data to model future outcomes (Javeed et al., 2023; Jdey et al., 2023; Pallathadka et al., 2023; Sahu et al., 2023). One example is using machine learning to predict and prevent forest fires by analyzing climate indicators, helping to make proactive decisions for environmental management (Alkhatib et al., 2023).

Many studies on the influence of climate conditions in Indonesia on indications of forest fires have been conducted by researchers. Nugrahani et al. (2024) used information on climate conditions (precipitation, dry spells, ENSO, and IOD) to predict the number of hotspots in Kalimantan as an indicator of forest fires by constructing an artificial neural network, random forest regression, gradient boosting, and Bayesian regression models. Mahendra et al. (2022) classified forest and land fires in Palembang, South Sumatra, using the C4.5 decision tree algorithm based on precipitation, wind speed, and air humidity information. Preeti et al. (2021) compared the decision tree algorithm, support vector machine, and random forest regression to predict forest fires based on information on climate conditions, including temperature, precipitation, wind, and air humidity.

Based on previous studies, many types of prediction models can be used to predict forest fires based on climate indicators, such as the Artificial Neural Network (ANN) model and the Random Forest model. ANN models have been widely used for predicting forest fires due to their ability to model complex, nonlinear relationships between climatic variables and fire occurrences (Jain et al., 2020). Random Forest models, on the other hand, offer robust prediction capabilities with high accuracy and are well-suited for handling large datasets (Kursa, 2014), making them ideal for predicting forest fires based on multiple environmental and climatic factors. The choice of an appropriate prediction model plays a critical role in forest fire prevention and mitigation, depending on the type of research being conducted and the specific fire indicators being used. If conducting research using a classification system such as Mahendra et al. (2022), then the prediction model is also a classification model. However, if researching to determine the effect of the relationship between climate indicators and the potential for forest fires, such as Nugrahani et al. (2024) and Preeti et al. (2021), then the regression model is more appropriate.

Hotspots are widely used as indicators of potential forest fires. Hotspots represent locations with a surface temperature above a certain threshold, identified through satellite imagery interpretation (Saharjo & Nasution, 2021). Generally, hotspots are spread randomly depending on the area's conditions, especially climate conditions. Hotspots are generally distributed across an area in a manner influenced by various environmental factors, particularly climate conditions. For instance,

areas with high precipitation tend to have fewer hotspots, while areas experiencing prolonged drought conditions are more likely to exhibit a higher concentration of hotspots (Giglio et al., 2018). The increasing number of hotspots correlates strongly with the likelihood of forest fires, as they often signal dry and combustible vegetation. Therefore, predicting the number of hotspots in a region based on climate variables, such as temperature or precipitation, could be a crucial step in forest fire prevention.

The number of hotspots based on climate conditions can be predicted by creating a model that can recognize the influence of the relationship between climate conditions and the number of hotspots so that the linear regression model is suitable. Currently, many regression models have been developed to overcome problems that have been experienced in the use of previous regression models. One of the developments of the regression model is the regularized regression model, which can overcome the problem of multicollinearity in data (Fikri et al., 2023; Herawati et al., 2018; Venkatesh et al., 2023). Therefore, this study uses regularized regression models, i.e., a regularization regression model consisting of ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO), and elastic-net as prediction models for the number of hotspots in Kalimantan based on climate indicators.

In addition to the regularization regression model, this study also uses a prediction model that applies the ensemble method, called Bayesian Model Averaging (BMA), based on a polynomial regression model. BMA is a model that bases its predictions on a weighted average of several models rather than just one model (Hinne et al., 2020). Both models were built using a combination of the best predictor variables based on the results of variable selection using the best subset selection method so that the resulting model does not have too high complexity and is easy to implement. Furthermore, the performance of both types of models was compared, and the best model was selected in predicting the number of hotspots in Kalimantan based on specific climate indicators, i.e., precipitation, precipitation anomalies, dry spells, El Nino Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD) indices, and seasonality.

This study aims to develop apredictive model for forecasting the number of hotspots in Kalimantan using a comprehensive statistical approach. The main objectives of this research are as follows. Selecting the Best Combination of Predictors: The first step in this study is to select the best combination of predictors for forecasting the number of hotspots in Kalimantan. This selection process will use the best subset selection method, with the criterion for selection being the lowest Bayesian Information Criterion (BIC) value. This approach ensures that the most relevant and informative predictors are chosen for the predictive model. Constructing Regularization Regression Models: After determining the set of predictors, several regularization regression models will be constructed, including ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO), and elastic-net regression. Additionally, a BMA model based on polynomial regression will also be developed. The use of these regularization methods aims to address multicollinearity and overfitting issues, ensuring that the models produced are stable and accurate. Determining the Best-Performing Model: The final step of the study is to evaluate the performance of each model in predicting the number of hotspots in Kalimantan. The evaluation will be based on two key metrics: Root Mean Square Error (RMSE) and the coefficient of determination (R^2), on both training and test data. The model with the best performance according to these metrics will be selected as the most effective model for accurate prediction. Through this systematic scientific approach, the study aims to produce a reliable predictive model that can be utilized to better understand and mitigate the risks of forest fires in Kalimantan.

This study contributes to the growing body of research on predictive modeling for forest fire management by addressing critical gaps in existing methodologies. While previous studies have utilized machine learning approaches such as Artificial Neural Networks (ANN) and Random Forest (RF) for forest fire prediction, these models often rely on single-model frameworks that may not adequately account for uncertainty or interactions among climate variables. In contrast, this research leverages BMA to integrate multiple model perspectives, enhancing robustness and predictive accuracy.

The novelty of this work lies in its application of BMA in combination with polynomial regression and best subset selection to predict forest fire hotspots. This approach allows for the inclusion of interaction terms and higher-order relationships among climate variables, such as dry spells, precipitation anomalies, and the Indian Ocean Dipole (IOD) index, which are often oversimplified in traditional regression models. Moreover, by incorporating uncertainty into predictions, the BMA model provides more reliable insights for decision-making, a feature critical for managing forest fires in complex environments like Kalimantan.

Additionally, this study emphasizes the role of climate variability indicators in fire hotspot prediction, offering a systematic methodology that bridges the gap between theoretical modeling and practical application. These contributions not only advance the predictive modeling field but also provide actionable insights for environmental management and policy formulation in tropical regions.

2. METHODS

2.1. Multiple Linear Regression

One of the simple and popular machine learning methods is linear regression. In principle, the linear regression method works by measuring the relationship between continuous variables, which, in this case, are assumed to have a linear relationship. The linear regression method that predicts a continuous variable based on more than one predictor variable is called multiple linear regression (Hope, 2020). The model produced by multiple linear regression is represented by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (1)$$

where y is a continuous response variable, β_0 is the intercept or intersection, which is defined as the value of the response variable when all predictor variables are zero, β_1 to β_n are the coefficients of the 1st to n-th predictor variables, and ε is the error of the model. Linear regression models involving polynomial variables such as x_i^2 or x_i^3 are called polynomial regression models (Montgomery et al., 2021).

According to Han et al. (2024), regression parameters β_i for $i = 0,1,2, \dots, n$ are estimated by minimizing the sum of the squared errors of the model represented by

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (2)$$

where y_i and $\hat{y}_i$ representing the actual and predicted values of the i-th observation point, respectively. Multiple linear regression is increasingly unable to work well using the ordinary least squares (OLS) as the number of predictor variables increases due to the increasing possibility of multicollinearity or linear relationships between predictor variables (Hope, 2020). Thus, a linear regression model with regularization such as ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO), and elastic-net have been developed which has an additional penalty to overcome multicollinearity.

2.2. Best Subset Selection

Best subset selection is a widely used variable selection method for selecting predictor variables in linear models. This method selects a combination of several predictor variables that produce the best model based on specific evaluation metrics such as BIC, adjusted R^2, and Mallows CP. Hastie et al. (2020) stated that if there is a vector Y of size $n \times 1$ containing the response variables, a matrix X of size $n \times k$ containing the predictor variables, and a subset of predictor variables with a size p between 0 and $\min\{n, k\}$, then the best subset selection method will find a combination of k predictor variables that produces the best model. The combination of predictor variables overcomes the problem in the context of squared errors, which are represented by

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^k} \|Y - \beta X\|_2^2 \quad (3)$$

where $\|\beta\|_0 \leq k$ dan $\|\beta\|_0 = \sum_{i=1}^{k} 1\{\beta_i \neq 0\}$.

The best subset selection method has advantages over other variable selection methods, such as forward selection and backward elimination. According to Brooks and Ruengvirayudh (2016), one of

the main advantages of best subset selection is choosing a model with the best combination of predictor variables by considering all models that can be formed based on the number of existing predictor variables. This advantage can overcome the limitations of the forward selection and backward elimination methods to produce a better model. However, if the number of predictor variables is enormous, this method must be considered due to the increasing computation time needed to create all possible models if the number of predictor variables increases (Brooks & Ruengvirayudh, 2016).

2.3. Regularized Regression Methods

Regularization is used to shrink the estimated value of the regression coefficient by providing a penalty when the model does not meet the multicollinearity assumption (Yanke et al., 2022). Regularization has three methods: ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO), and elastic-net.

2.3.1. Ridge Regression

The first regularization method is ridge regression, introduced by Hoerl (1962). Ridge regression overcomes multicollinearity by determining a biased estimator but has a smaller variance value than the variance value in multiple linear regression (Wasilaine et al., 2014). Meanwhile, multicollinearity is a problem that occurs due to two or more correlated predictor variables. According to Saleh et al. (2019), ridge regression provides a penalty to the model to provide limits for the coefficient values of the linear regression model so that they do not have tremendous values without limits.

Ridge regression is represented by

$$\hat{\beta} = \arg\min_{\beta} (Y - X\beta)^T (Y - X\beta) + \lambda \|\beta\|_2^2 \tag{4}$$

where $\beta_{k\times 1}$ contains the regression coefficients to be estimated, $Y_{n\times 1}$ contains the response variables, $X_{n\times k}$ contains the predictor variables, λ is a shrinkage parameter whose value is always positive, and $\|\beta\|_2^2$ is a ridge penalty whose value is equal to $\sum_{j=1}^{k} \beta_j^2$ (Saleh et al., 2019). If λ approaches zero, the value of the regression coefficient will be greater as in OLS regression, but if $\lambda \to \infty$ then the value of the coefficient will be closer to zero.

2.3.2. Least Absolute Shrinkage and Selection Operator (LASSO) Regression

Ridge regression has a disadvantage, i.e., it can only shrink the regression coefficient to near zero, so a regularization method was introduced to overcome this deficiency. Tibshirani (1996) first introduced the Least Absolute Shrinkage and Selection Operator (LASSO) method, which was used to overcome the multicollinearity problem (Andana et al., 2017). According to Saleh et al. (2019), LASSO can overcome the shortcomings of ridge regression by shrinking the regression coefficient to zero. Therefore, LASSO is suitable for high-dimensional data because it can reduce the predictor variables used. If there is high-dimensional data that has as many predictor variables as k and the number of data as n where $k > n$, then LASSO can be represented by

$$\hat{\beta} = \arg\min_{\beta} (Y - X\beta)^T (Y - X\beta) + \lambda \|\beta\|_1 \tag{5}$$

where $\|\beta\|_1$ represents the LASSO penalty, which has the same value as $\sum_{j=1}^{k} |\beta_j|$.

2.3.3. Elastic-Net Regression

Zou and Hastie (2005) introduced a combined method of ridge regression and LASSO, i.e., elastic-net regression. According to Handayani and Wachidah (2022), the advantages of this method are that it can handle multicollinearity problems, can reduce the regression coefficient to precisely zero, can select predictor variables from a group of correlated predictor variables, and can select variables simultaneously. Elastic-net regression can be represented by

$$\hat{\beta} = \arg\min_{\beta} (Y - X\beta)^T (Y - X\beta) + \lambda [\alpha \|\beta\|_1 + (\alpha - 1)\|\beta\|_2^2] \tag{6}$$

where $[\alpha \|\beta\|_1 + (\alpha - 1)\|\beta\|_2^2]$ is the elastic net penalty with $\alpha \in [0,1]$. If $\alpha = 0$, Eq. 6 becomes the same as the ridge regression equation, while when $\alpha = 1$, Eq. 6 becomes the same as the LASSO regression equation.

2.4. Bayesian Model Averaging (BMA)

BMA is a model that bases its predictions on a weighted average of several models rather than just one model (Alhassan et al., 2024; Huang et al., 2023). BMA applies an ensemble method by combining multiple models based on each model's posterior probability or weight. According to Claeskens and Hjort (2008), if there is a collection of models $M_1, \ldots, M_m$ that predict the value of the response variable y from data D, then BMA bases its prediction results not only using one model but combining all models based on their posterior probabilities. BMA is represented by

$$P(y|D) = \sum_{k=1}^{m} P(M_k|D)P(y|M_k, D) \tag{7}$$

where $P(y|D)$ is the weighted average of the posterior densities of y given the data D, $P(M_k|D)$ is the posterior probability of the model M_k given the data D, and $P(y|M_k, D)$ is the posterior density of y, when the model M_k is the most appropriate model.

$P(M_k|D)$ or the posterior probability of the model M_k is obtained by applying Bayes' theorem and is represented by

$$P(M_k|D) = \frac{P(M_k)\lambda_{n,k}(D)}{\sum_j^m P(M_j)\,\lambda_{n,k}(D)} \tag{8}$$

where $P(M_k)$ is the prior probability of the model M_k, which is typically distributed uniformly $1/m$, while $\lambda_{n,k}(D)$ is the marginal density of the data D represented by

$$\lambda_{n,k}(D) = \int L(D, \theta_k)P(\theta_k|M_k)\, d\theta_k \tag{9}$$

where θ_k is a vector of the parameters of the model M_k, $L(D, \theta_k)$ is the likelihood function of the model M_k, and $P(\theta_k|M_k)$ is the prior density of the model M_k (Claeskens & Hjort, 2008).

Just as the BMA model equation is a weighted average of the posterior densities y, the posterior mean value of the BMA model, denoted by E, is a weighted average of the posterior mean values for each model M_k and is represented by

$$E(y|D) = \sum_{k=1}^{m} P(M_k|D)E(y|M_k, D) \tag{10}$$

where $E(y|M_k, D)$ is the posterior mean value when the model M_k is the most appropriate model.

2.5. Bayesian Information Criterion (BIC)

Various criteria in variable selection have been developed and widely used in various studies. According to Dziak et al. (2020), several criteria in variable selection can be defined as a log-likelihood function with a penalty known as the Information Criterion (ICs). The main objective of ICs is to select a model that minimizes the value of

$$IC = -2l + C_n p \tag{11}$$

where l is the log-likelihood function of the model, C_n is a constant or penalty function whose type depends on the ICs criteria used, n is the number of sample data, and p is the number of parameters in the model (Dziak et al., 2020).

One type of IC widely used and has been widely developed is the Bayesian Information Criterion (BIC). BIC is a criterion widely used in variable selection methods and focuses on models with low complexity. BIC works by giving a high penalty to models with high complexity represented by the number of parameters in the model so that it can reduce the potential for overfitting (Kasali & Adeyemi, 2022). The penalty used in BIC, represented by C_n, is the $\ln(n)$ function (Dziak et al., 2020). The BIC differs from other types of ICs, such as the Akaike Information Criterion (AIC), which focuses on models with high accuracy, so selecting these two criteria is based on research needs.

2.6. Evaluation Metrics

Evaluation metrics are measurement criteria that are often used to determine the level of accuracy and feasibility of a model. One widely used evaluation metric is Root Mean Squared Error (RMSE).

RMSE is a method used to measure a prediction model's accuracy level as a form of model evaluation (Sanjaya & Heksaputra, 2020). RMSE calculates the average squared value of the number of errors in a prediction model. The lower the RMSE value, the more accurate the prediction model produced. The RMSE value is calculated by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{12}$$

where n represents the number of data, y_i and $\hat{y}_i$ represents the actual and predicted value of the response variable from the i-th observation.

In addition to RMSE, another widely used evaluation metric is the coefficient of determination or R-squared (R^2). R^2 is one of the widely used model evaluation metrics to measure the performance of a prediction model. The R^2 value is used to measure the variance in the response variable that can be explained by the model (Purwanto & Sudargini, 2021). The R^2 value ranges from 0 to 1, with the higher the value, the better the model (Chicco et al., 2021).

3. STUDY AREA AND DATASETS

3.1. The Island of Borneo

Borneo, the third-largest island in the world following Greenland and New Guinea, is a critical region for biodiversity and environmental science (Keong & Onuma, 2021). The island of Borneo covers an area located between 4°S-7°N and 108°E-120°E, covering approximately 743,330 km^2 area (Sa'adi et al., 2020). The island, situated in Southeast Asia, is divided among three countries: Indonesia, Malaysia, and Brunei. The tropical equatorial climate of Borneo is categorized as tropical rainforest (Af), with uniform temperature all year round. Precipitation is substantial, averaging over 3000 mm annually, and is distributed throughout the year, contributing to the island's lush rainforests. Borneo represents a critical region for both scientific inquiry and conservation efforts. Its rich biodiversity, unique geological features, and significant environmental challenges make it a focal point for research to understand and preserve one of the world's most important natural areas (Keong & Onuma, 2021; von Rintelen et al., 2017).

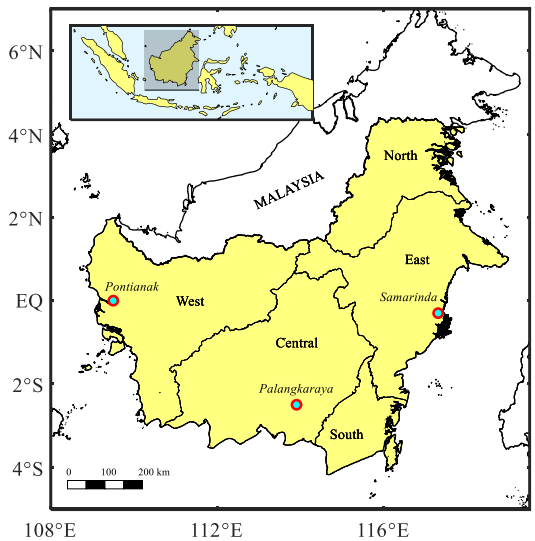


Fig. 1. Map of the island of Borneo.

Kalimantan is the Indonesian portion of the island of Borneo, covering roughly 73% of the island's land area. Kalimantan encompasses five provinces: West, Central, South, East, and North Kalimantan, as shown in **Figure 1**. Each province has distinct geographical features, from the coastal

plains of West Kalimantan to the rugged highlands of East Kalimantan. Kalimantan is highly prone to fires due to natural and human-induced factors. These fires, often called "forest fires" or "slash-and-burn" fires, have significant environmental, social, and health impacts (Harrison et al., 2024). Fires contribute to significant deforestation, loss of biodiversity, and degradation of ecosystems. Peatland fires release large amounts of carbon dioxide and other greenhouse gases, exacerbating climate change (Palamba, 2024). Furthermore, smoke from fires causes severe air pollution, leading to respiratory problems and other health issues for local communities and even affecting neighbouring countries (Sambodo et al., 2024). Moreover, fires can damage crops, disrupt livelihoods, and incur significant costs for firefighting and restoration efforts.

In the lush landscapes of Kalimantan, the natural causes of forest fires are an integral aspect of the region's ecological dynamics. Forest fires in Kalimantan can occur due to several natural factors, each contributing to the intricate fire ecology of this tropical environment. The natural factors contributing to forest fires in Kalimantan are diverse and interconnected. Lightning strikes, land cover, spontaneous combustion, and seasonal weather patterns all play roles in the fire ecology of this region (Barros et al., 2021; Edwards et al., 2020). Understanding these factors is crucial for managing and mitigating the impact of forest fires in Kalimantan. Natural variations in weather patterns, such as El Niño events, can significantly impact fire risks (Brasika et al., 2021; Nurdiati, et al., 2022a). During El Niño years, Kalimantan often experiences prolonged dry periods, reducing soil moisture and increasing the flammability of vegetation. These weather patterns create ideal conditions for fires to ignite and spread, whether from natural or anthropogenic sources.

3.2. Sources and Types of Datasets

This study used hotspot data as an indicator of forest fire in Kalimantan. Hotspots are a crucial indicator for monitoring and predicting forest fires in Kalimantan, as well as in other fire-prone regions (Kadir et al., 2023; Usup & Hayasaka, 2023). In the context of forest fires, a hotspot refers to a location with an unusually high surface temperature, which can be detected using remote sensing technologies. These hotspots indicate areas where combustion or intense heating occurs, often associated with fire activity. In Kalimantan, these hotspots typically emerge during drought, exacerbated by various climatic phenomena.

Total precipitation is one of the factors that influences fire activity (Fanin & Van Der Werf, 2017). During months with below-average precipitation, the forest biomass becomes drier and more susceptible to ignition. In Kalimantan, the relationship between precipitation and fire hotspots is inversely correlated; the likelihood of fire hotspots increases as precipitation decreases. Conversely, higher precipitation levels dampen the forest floor, reducing fire susceptibility. In addition to total precipitation, precipitation anomalies are a climate factor that can significantly impact fire risk (Nurdiati et al., 2022b). Precipitation anomalies refer to deviations from normal precipitation patterns. During years of significant precipitation deficits (often linked to broader climatic trends), Kalimantan experiences an uptick in fire hotspots. For instance, a negative precipitation anomaly can lead to prolonged dry spells, creating ideal conditions for fire ignition and spread. Monitoring these anomalies helps predict potential fire outbreaks, enabling timely intervention.

The number of dry days, called dry spells, is another critical factor in fire dynamics (Kumar & Kumar, 2022; Najib et al., 2024). In Kalimantan, prolonged dry spells can dry out surface litter and deeper soil moisture, making the region more susceptible to fire. Statistical analyses have shown that fire hotspots are more prevalent during periods exceeding a certain threshold of dry spells, emphasizing the cumulative effect of dryness.

The El Niño Southern Oscillation (ENSO) significantly impacts global weather patterns, including Kalimantan. During El Niño years, the region often experiences drier-than-normal conditions, increasing fire hotspots (Najib et al., 2022a). The resulting drought stress on vegetation heightens the risk of fires as lower humidity and higher temperatures prevail. Conversely, during La Niña events, increased precipitation typically reduces fire occurrences. Understanding ENSO patterns aids in predicting fire risk and implementing pre-emptive measures. Elsewhere, the Indian Ocean Dipole (IOD) also plays a vital role in influencing precipitation patterns in Kalimantan. A positive

IOD phase often correlates with drier conditions, further exacerbating the risk of fire hotspots (Nurdiati et al., 2022a). Conversely, a negative IOD phase typically brings increased precipitation, which can mitigate fire risk. The interplay between IOD phases and local climate conditions underscores the complexity of fire dynamics in the region. Moreover, seasonality is another critical factor in understanding fire hotspots in Kalimantan. The dry season, particularly from July to September, often sees the highest incidence of fires (Najib et al., 2022b), coinciding with lower precipitation and higher temperatures. Human activities, such as land clearing for agriculture, often peak during this time, further increasing the risk of ignition.

The interplay of natural factors creates a complex web of influences on fire dynamics. The relationship between fire hotspots in Kalimantan and natural factors like precipitation, climatic indices, and seasonality is intricate and multifaceted. Understanding the interactions between these factors is essential for effective fire management and conserving this vital ecosystem. In this study, we used multiple sources of datasets for each variable ranging from January 2001 until December 2020. The hotspot data comes from the Indonesian Agency for Meteorological, Climatological, and Geophysics, which is processed data from MODIS sensors of the Terra and Aqua satellites curated to exclude false fire hotspots. Meanwhile, local climate data is obtained from CMORPH-CRT, and global climate data is sourced from PSL NOAA. For more details, **Table 1** briefly describes the sources and types of data used.

Table 1.

Description of the sources and types of data used.

Variable	Resolution	Source
Hotspots (Y)	Monthly, 0.25 × 0.25	Retrieved from Indonesian Agency for Meteorological, Climatological and Geophysics
Total Precipitation (X_1)	Monthly, 0.25 × 0.25	Retrieved from monthly data of CMORPH_CRT datasets https://ftp.cpc.ncep.noaa.gov/precip/PORT/SEMDP/CMORPH_CRT/DATA/
Precipitation Anomalies (X_2)	Monthly, 0.25 × 0.25	Retrieved from monthly data of CMORPH_CRT datasets https://ftp.cpc.ncep.noaa.gov/precip/PORT/SEMDP/CMORPH_CRT/DATA/
Dry spells (X_3)	Monthly, 0.25 × 0.25	Retrieved and processed from daily data of CMORPH_CRT datasets https://ftp.cpc.ncep.noaa.gov/precip/PORT/SEMDP/CMORPH_CRT/DATA/
ENSO index (X_4)	Monthly, time series	Retrieved from https://psl.noaa.gov/gcos_wgsp/Timeseries/Nino34/
IOD index (X_5)	Monthly, time series	Retrieved from https://psl.noaa.gov/gcos_wgsp/Timeseries/DMI/
Seasonality (X_6)	Monthly	Month of data

Based on previous studies (Najib et al., 2021), it is well-established that precipitation predictors derived from meteorological data significantly influence the occurrence and distribution of forest fire hotspots in Kalimantan. These predictors capture the dynamic relationship between rainfall patterns and fire activity, helping them to understand and potentially mitigate fire risks. Among these, three predictors stand out due to their strong correlation with hotspot frequency: the two-month average precipitation, the monthly precipitation anomaly, and the three-month number of dry days.

The two-month average precipitation, denoted as X_1, represents the mean rainfall over the observation month and the preceding month. This predictor provides a broader temporal context, smoothing out short-term fluctuations and highlighting the cumulative precipitation available to reduce fire risk. Lower values of X_1 are generally associated with drier conditions that can exacerbate forest fire risks by reducing soil moisture and vegetation dampness.

The monthly precipitation anomaly, referred to as X_2, quantifies the deviation of rainfall during the observation month compared to its long-term historical average, or "normal". This measure serves as a critical indicator of abnormal climatic conditions, such as extended dry spells or unusually wet periods, which can influence fire susceptibility. To ensure clarity in this study, the term "precipitation anomaly" is defined inversely to the typical interpretation: a positive anomaly value signifies a deficit in rainfall relative to the norm, while a negative value indicates an excess of precipitation. This reversed convention is intentional, facilitating a direct association between positive X_2 values and increased fire risks due to insufficient rainfall.

Lastly, the number of dry days over a three-month period, labeled as X_3, captures the cumulative count of days with minimal precipitation, defined as less than 1 mm of rainfall per day. This predictor spans the observation month and the two preceding months, reflecting extended periods of dryness that are crucial for understanding fire dynamics. Prolonged dry periods, as indicated by higher X_3 values, often lead to reduced soil moisture and increased flammability of vegetation, creating favorable conditions for forest fires.

Together, these derived precipitation predictors form a comprehensive framework for analyzing the influence of rainfall variability on forest fire hotspots. By capturing both temporal trends and deviations from the norm, they provide valuable insights into the climatic drivers of fire activity in Kalimantan, enabling more effective risk assessment and management strategies.

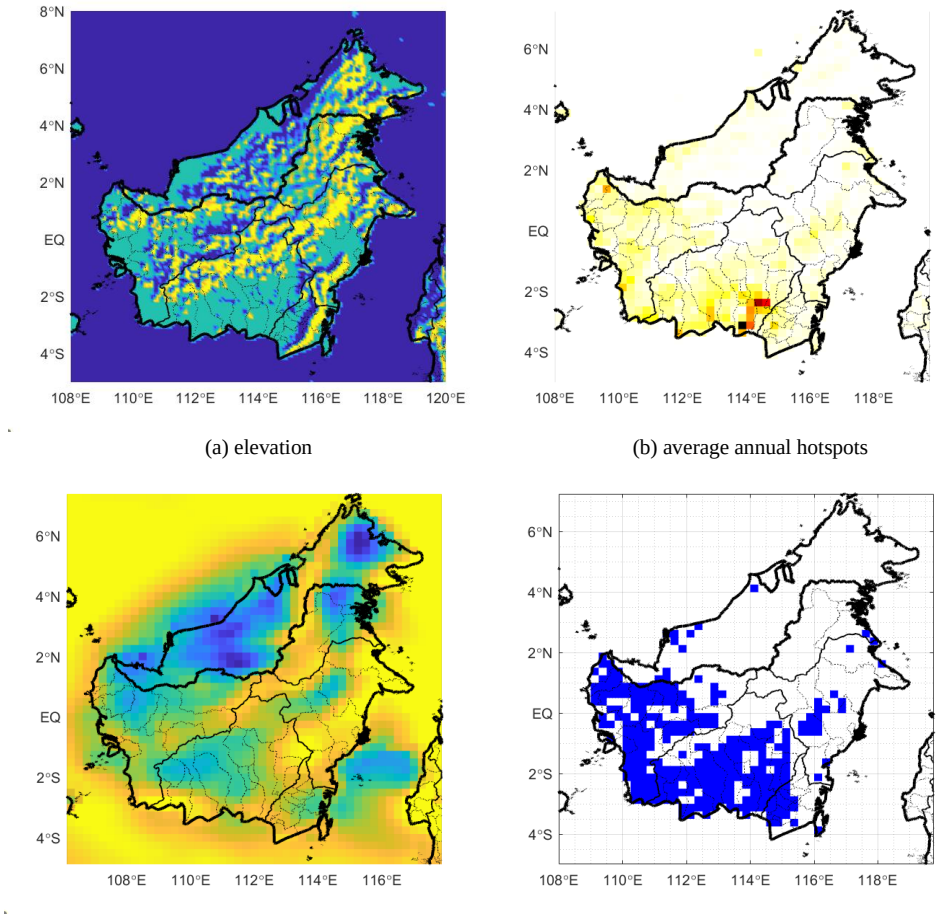
4. RESULTS AND DISCUSSION

4.1. Pre-processing of Datasets

In our quest to understand the dynamics of fire hotspots in Kalimantan, we begin by processing satellite data in **Table 1**. This approach allows us to pinpoint specific grid points that are significant to our study of fire incidents. Our research focuses primarily on lowland regions, such as Central and Western Kalimantan, where fire occurrences are more frequent (**Figure 2a** and **2b**). These areas offer critical insights for effective fire management and climate studies. In contrast, regions with high precipitation, such as the Malaysian part of Borneo, experience fewer fire hotspots and are therefore not central to our analysis (**Figure 2c**). The abundant precipitation in these areas complicates the correlation between climate data and fire events, so our research prioritizes the more fire-prone lowland regions.

To focus on regions significantly impacted by fire incidents, we applied a k-means clustering algorithm to identify grid points with the highest correlations to hotspot occurrences. This clustering method grouped spatial data based on similarity in fire activity and climatic conditions, allowing us to isolate areas most vulnerable to fire risks. **Figure 2d** illustrates the clustering results, highlighting the critical grid points selected for further analysis. These areas, primarily in lowland regions of Central and Western Kalimantan, were used to aggregate data for temporal modeling. By employing k-means clustering, we ensured that the analysis targeted regions with consistent patterns of fire occurrence, optimizing the model's ability to capture the relationship between climate indicators and hotspots. This spatial pre-selection process reduced noise in the dataset caused by areas with low fire activity or high precipitation, such as the Malaysian part of Borneo. The selected grid points were aggregated into a time series format, enabling the study to focus on temporal dynamics while retaining spatial relevance through targeted grid selection.

By aggregating the relevant grid points into a time series, we can effectively track fire occurrences over time, identifying patterns and trends that emerge in relation to climatic variations. This focused methodology allows us to disentangle the complex relationships between precipitation and fire activity. Our satellite data processing aims to illuminate the region’s most vulnerable to fire risks while acknowledging the intricate interplay of climatic factors. By concentrating on the low-precipitation areas with high-fire incidents, we hope to provide valuable insights that inform targeted interventions and enhance our understanding of fire dynamics in Kalimantan's unique environment.



(a) elevation (b) average annual hotspots

(c) average annual precipitation (d) selected area using k-mean clustering

Fig. 2. Map of the island of Borneo showing: a) elevation, b) average annual hotspots, c) average annual precipitation, and d) selected area using k-mean clustering.

4.2. Correlation Between Variables

In this section, we conduct a cross-correlation analysis between fire hotspots and climatic factors, as shown in **Figure 3**. Our analysis reveals that the highest correlation with fire hotspots is observed with the variable dry spells, yielding a correlation coefficient of 0.699. This strong positive correlation suggests that prolonged dry conditions significantly increase fire occurrences. Additionally, we find the highest correlation among the climatic factors is between total precipitation and dry spells, with a coefficient of -0.876. This inverse relationship indicates that as total precipitation increases, the frequency or duration of dry spells tends to decrease, which aligns with the understanding that wetter conditions typically mitigate fire risk.

Given these correlations, the implications for predicting fire hotspots using climatic factors are substantial. The strong correlation between dry spells and fire hotspots implies that monitoring and forecasting dry conditions could effectively predict fire risks. In regions with anticipated dry spells, proactive measures can be implemented to mitigate potential fire outbreaks. Furthermore, the inverse relationship between total precipitation and dry spells emphasizes the importance of rainfall patterns in fire prediction models. By integrating total precipitation and dry spells into predictive models, we can improve the accuracy of forecasts regarding fire hotspots. Additionally, understanding the interactions between these climatic factors—such as the influence of the ENSO and IOD indices—can provide further insights into seasonal variations and long-term trends in fire occurrences.

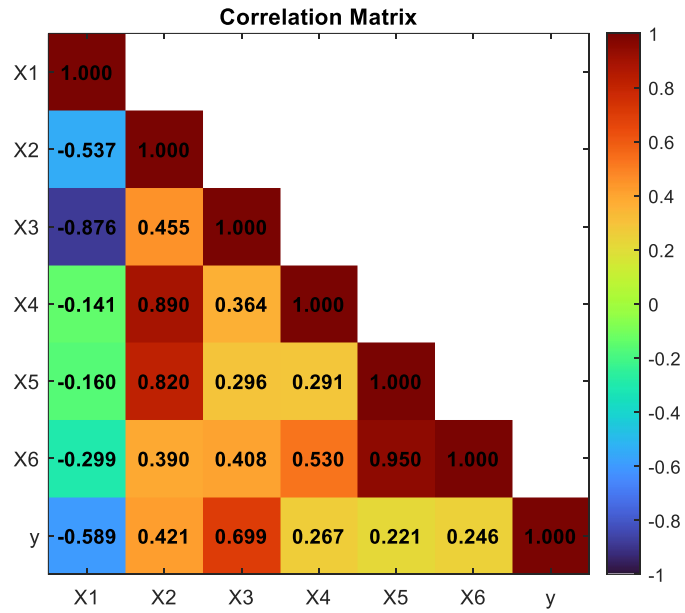


Fig. 3. Heatmap of cross-correlation analysis between fire hotspots and climatic factors.

4.3. Polynomial Features and Variable Selection

The weak linear relationship observed between the response variable and the predictor variable, as shown in **Figure 3**, indicates that the data may not be suitable for input for a linear regression model. This weak correlation can lead to a phenomenon known as underfitting, where the model fails to learn and represent the underlying patterns in the data adequately. Consequently, adopting alternative methods that enhance the model's ability to capture these subtle relationships becomes essential. One such method is the application of polynomial features, as proposed by Maulud and Abdulazeez (2020). This technique transforms the predictor variables by applying specific polynomial degrees, thus generating additional predictor variables. The primary objective of this approach is to increase the complexity of the model, enabling it to learn better and fit the weak linear relationship between the response and predictor variables. The model can address the challenges posed by underfitting by employing polynomial features, improving its predictive performance, and providing a more accurate data representation.

In addition to the challenges of weak linear relationships, linear regression models struggle to capture the interaction effects among variables directly. This is particularly relevant in the context of this study, where the predictor variables consist of multiple climate indicators that typically interact with one another. Adding additional predictor variables in the form of interaction terms becomes necessary to address this limitation. According to Bertsimas and Wiberg (2020), incorporating interaction variables can significantly enhance the model's ability to identify and learn from weak linear relationships among the variables.

Considering this, this research integrates interaction terms as supplementary predictor variables, aiming to effectively capture the interactions between climate indicators in predicting the number of hotspots while simultaneously improving the model's accuracy. The addition of these interaction variables is implemented through the polynomial features method. By applying polynomial transformations up to a maximum degree of three, the analysis generates a total of seventy-seven additional predictor variables, as detailed in **Table 2**. This strategic enhancement broadens the scope of the model's input and allows for a more nuanced understanding of the interplay among climate indicators, leading to more precise and reliable predictions.

Table 2.

Additional predictor variables resulting from the application of polynomial features.

Type of Predictor Variable	Number of Predictors	Members of predictors
variables of degree 2	6	$x_1^2, x_2^2, \dots, x_6^2$
variables of degree 3	6	$x_1^3, x_2^3, \dots, x_6^3$
interaction variables	65	$x_1x_2, x_1^2x_2, \dots, x_1x_5x_6$

The inclusion of additional predictor variables, as illustrated in **Table 2**, brings the total number of predictor variables to eighty-three. This substantial increase in variable count can significantly prolong the computational time required to develop an accurate predictive model. To address this issue, a variable selection technique is employed to reduce the number of predictor variables, streamlining the modelling process. The dataset must first be divided into training and testing subsets before applying the variable selection technique to ensure a valid and reliable model evaluation. The training data, covering the period from January 2001 to December 2018, is utilized for model development, while the testing data, spanning from January 2019 to December 2020, assesses the model's performance.

The selection of these specific time ranges is based on a series of experiments that explored various proportions of data allocation. This approach led to the identification of the optimal division that effectively balances the training and testing datasets. Ensuring that the training and testing periods do not overlap, the model is evaluated on data it never encountered during training. This methodology is essential for validating the robustness of the model's performance, as it allows for a more accurate assessment of how well the model can generalize to unseen data. This strategic division enhances the credibility of the results and contributes to a more reliable predictive framework.

Subsequently, all predictor variables in the training dataset are selected using the best subset selection technique. This approach employs the Bayesian Information Criterion (BIC) as the metric for determining the optimal combination of predictor variables. A lower BIC value indicates a better model fit, which allows us to identify the most effective combination of predictor variables for inclusion in the final model. The maximum number of predictor variables selected through this technique is limited to ten. The results of the variable selection process using best subset selection are visualized in a plot, as shown in **Figure 4**. This visualization clearly represents the selected variables and their respective contributions to the model, facilitating a better understanding of the relationships within the data. By employing this rigorous selection method, the model can achieve a more refined and efficient representation of the underlying patterns, enhancing its predictive capabilities.

Figure 4 illustrates the ten combinations of predictor variables and their corresponding BIC values. Each line in the plot indicates the selected predictor variables for each combination, with the line length reflecting the order in which the variables were chosen. For instance, the combination represented in the second row from the bottom, which yields a BIC value of -293.5739, includes the variables x_3^3 and x_3^2.

The combinations with the highest BIC values appear at the bottom of the plot, while those with the lowest values are positioned at the top. Consequently, the model that results in the lowest BIC value is constructed from a combination of six predictor variables, as highlighted in **Figure 4** and presented in Eq. 13. This model achieves a BIC value of -307.6923, marking it as the best result after conducting various iterations of the best subset selection process with different maximum numbers of predictor variables.

The results from the best subset selection, indicated by Eq. 13, revealed a model comprised of six terms in polynomial form, highlighting the interactions between precipitation anomalies (x_2), dry spells (x_3), and the Indian Ocean Dipole (IOD, x_5) index. This model illustrates the complex relationships among these variables and their combined effect on hotspot occurrences. This finding suggests that these three climate indicators strongly influence the number of hotspots in Kalimantan, underscoring their critical role in environmental monitoring and predictive modelling.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_3^3 + \hat{\beta}_2 x_3^2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_2 x_3^2 + \hat{\beta}_5 x_3^2 x_5 + \hat{\beta}_6 x_2^2 x_5 \quad (13)$$

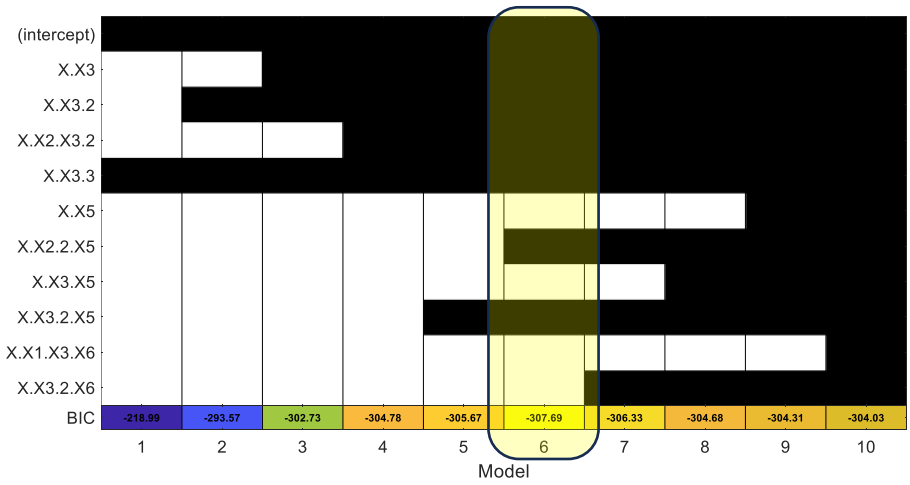


Fig. 4. Results of variable selection using best subset selection.

A key takeaway from the analysis is the prominent role of dry spells as a significant predictor for hotspots. Their presence in nearly every term of the equation underscores their critical influence on environmental conditions that contribute to hotspot formation. Dry spells can exacerbate drought conditions, reducing soil moisture and increasing the likelihood of fire events, leading to higher hotspot counts. The interactions identified in the model further emphasize the importance of considering how these factors interrelate. For instance, the impact of precipitation anomalies may be modulated by the presence of dry spells, indicating that an increase in precipitation does not necessarily mitigate hotspots if dry conditions persist concurrently.

This comprehensive understanding enhances predictive accuracy and provides valuable insights for stakeholders involved in environmental monitoring and management. By recognizing the interplay between these climatic factors, strategies can be developed to address and mitigate the risks associated with hotspot occurrences, particularly in vulnerable regions.

4.4. Training for regularized regression and BMA models

The prepared data was subsequently utilized to develop a regression model based on Equation (13) to predict the number of hotspots in Kalimantan. The first phase of model development involved training the model using a training dataset. This training process encompassed several regression models, specifically regularized regression techniques, including ridge, LASSO, and elastic-net regressions, alongside a BMA approach based on polynomial regression.

4.4.1. Regularized Regression Models

The training of the three regularized regression models begins with determining the hyperparameter values using a 10-fold cross-validation method. This approach is employed to identify the best lambda values for the ridge and LASSO regression models, while alpha and lambda hyperparameters are assessed for the elastic-net regression model. For the ridge and LASSO models, a total of 100 lambda values are randomly sampled from the range 10^{-10} to 10^{10} for application in the cross-validation process. In addition, the alpha value for the elastic-net model is determined by randomly selecting 20 alpha values from 0 to 1. During this procedure, the best lambda is applied in conjunction with each trial of alpha during cross-validation.

The selection of the best hyperparameters is based on the lowest Mean Squared Error (MSE) observed during the validation process. The resulting hyperparameter values for alpha and lambda for each regression model, derived from cross-validation, are summarized in **Table 3**. This systematic approach ensures the models are finely tuned for optimal performance, facilitating more accurate predictions in subsequent analyses.

Table 3.

Optimal hyperparameter values used in each regularized regression model.

Regularized regression model	alpha	Lambda
Ridge	0*	0,01917910
LASSO	1*	0,03053856
Elastic-net	0,6841737	0,17868010

**This hyperparameter value is a fixed value.*

The hyperparameter alpha values for the ridge and LASSO regression models, as indicated in **Table 3**, are fixed at 0 and 1, respectively. This specific allocation signifies the contrasting nature of these two regression techniques, with ridge regression favoring the inclusion of all predictors while LASSO regression actively eliminates some, leading to sparse models. In contrast, the elastic-net regression model employs an alpha value of 0.68, which is notably closer to 1, indicating a predominant reliance on the LASSO mechanism. This suggests that while elastic-net incorporates both ridge and LASSO elements, its behavior is more aligned with LASSO, thus promoting sparsity in the coefficient estimates.

Moreover, the lambda values reflect the degree of regularization imposed on the coefficients of each regression model. The elastic-net model exhibits the highest lambda value of 0.18, as presented in **Table 3**, which is significantly greater than the lambda values of the ridge and LASSO models, both of which remain under 0.05. This elevated lambda in the elastic-net model results in a more substantial penalty on the coefficients, driving them closer to zero compared to those in the ridge and LASSO models. Consequently, this pronounced regularization effect of elastic-net can lead to better performance in scenarios where multicollinearity is present or when there are many predictors, as it combines the strengths of both regularization techniques while managing to reduce overfitting effectively. The interplay of these hyperparameters ultimately illustrates the nuanced approach of elastic-net regression, making it a powerful tool for tackling complex predictive modeling tasks.

After determining the optimal hyperparameter values for alpha and lambda for each model, as presented in **Table 3**, these values were subsequently utilized to train the three regularized regression models for predicting the number of hotspots in Kalimantan. The resulting models included ridge regression, LASSO, and elastic-net, each characterized by distinct intercepts and coefficients detailed in **Table 4**.

Table 4.

Intercept and coefficient values for each predictor variable in the three regularized regression models.

Coeficients*	Regularized regression models		
	Ridge	***LASSO***	***Elastic-net***
$\boldsymbol{\beta_0}$	-3556,39	-4110,76	-1888,26
$\boldsymbol{\beta_1}$	0,122	0,129	0,099
$\boldsymbol{\beta_2}$	-11,50	-12,50	-8,51
$\boldsymbol{\beta_3}$	356,33	398,11	230,76
$\boldsymbol{\beta_4}$	0,043	0,042	0,046
$\boldsymbol{\beta_5}$	-0,384	-0,386	-0,377
$\boldsymbol{\beta_6}$	107,59	106,21	111,31

**Intercept and coefficient values for the best variable combination based on the results of best subset selection in equation (13).*

The results of the regularized regression analysis indicate that the absolute value of the highest coefficient is attributed to the interception. This suggests a strong baseline level of prediction for the model when all predictors are held constant at zero.

Among the predictor coefficients, the highest positive values are observed for β_3 and β_6. Specifically, β_3 representing the coefficient for dry-spells, underscores its significant influence on the prediction of hotspots. This reinforces the finding that dry spells are critical factors contributing to fire risks, as they can create conditions conducive to ignition and spread.

In addition, β_6 which captures the interaction between precipitation anomalies and the IOD index, highlighting the importance of understanding how these two factors work together. This interaction suggests that changes in precipitation patterns, influenced by IOD, can exacerbate or mitigate the effects of dry spells on hotspot occurrences.

4.4.2. Bayesian Model Averaging (BMA) Model

The training of the BMA model, based on polynomial regression, begins by determining the parameters of the BMA model, which consists of the coefficients and weights of the polynomial regression model. Determining the polynomial regression model coefficients is done by generating all possible polynomial regression models based on various combinations of predictor variables and estimating the regression coefficients for each model. Six polynomial regression models are then selected based on the lowest BIC (Bayesian Information Criterion) values, after which each model is assigned, a weight representing the influence of the selected models on the final BMA model. **Table 5** displays the interceptive values, regression coefficients, and weights for each selected polynomial regression model, obtained from the BMA model training process.

Table 5.

Summary of the BMA model components along with their respective weights.

Coefficients	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$\boldsymbol{\beta_0}$	-5426,87	-6250,85	-5853,77	-5258,17	-6764,41	-6181,12
$\boldsymbol{\beta_1}$	0,148	0,166	0,157	0,144	0,178	0,165
$\boldsymbol{\beta_2}$	-14,88	-16,88	-15,98	-14,55	-18,20	-16,80
$\boldsymbol{\beta_3}$	497,45	567,61	537,13	485,77	614,88	564,54
$\boldsymbol{\beta_4}$	0,039	0,00	0,042	0,044	0,00	0,00
$\boldsymbol{\beta_5}$	-0,391	-0,411	-0,214	0,00	-0,225	0,00
$\boldsymbol{\beta_6}$	103,05	109,05	0,00	0,00	0,00	0,00
Weight	**0,405**	**0,254**	**0,147**	**0,094**	**0,066**	**0,034**

Table 5 shows that each selected polynomial regression model has a different number of predictor variables. Model 1 is the model that uses all the predictor variables and has the most predictors, while Model 6 is the model with the fewest predictor variables, using only 3. Additionally, it can be observed that the intercept and regression coefficient values for each selected model tend to be small, with some coefficients approaching zero, similar to the regularization regression model. **Table 5** also displays the weights for each selected model, where the total sum of the weights equals 1. Model 1 has the most significant weight, with six predictor variables, indicating that this model best explains the overall data. Therefore, Model 1 will influence the final BMA model the most. Meanwhile, Model 6, with 3 predictor variables, has the most negligible weight, even below 5%.

All the selected polynomial regression models will be combined into a single final BMA model based on the weights shown in **Table 5**. **Table 6** provides detailed insights into the coefficients β of the Bayesian regression model, describing the relationship between each predictor and the response variable. The coefficient column represents the expected effect of each predictor on the response variable. For instance, β_3 has a coefficient of 530, suggesting that a unit increase in the corresponding predictor is associated with an average increase of 530 units in the response variable. These coefficients provide critical insights into the strength and direction of the relationships in the model.

The $p \neq 0$ column indicates the posterior probability that each coefficient is not zero. Predictors β_1, β_2, and β_3 have $p \neq 0$ values of 100%, reflecting strong evidence of their significant impact on the response variable. On the other hand, predictors such as β_4 (64.6%) and β_6 (65.9%) exhibit lower probabilities, suggesting weaker evidence of their relevance. These probabilities highlight the varying levels of certainty about the predictors' importance within the model. Overall, **Table 6** identifies β_1, β_2, and β_3 as the most significant predictors due to their high posterior probabilities, making them reliable contributors to the model. In contrast, β_4, β_5, and β_6 show weaker evidence for inclusion, with lower probabilities, indicating higher uncertainty about their effects.

Table 6.

Summary of the Bayesian Model Averaging (BMA) model.

Coefficient	p!=0 (%)	EV	SD	Upper PI (95%)	Lower PI (95%)
$\boldsymbol{\beta_0}$	100	-5797	2038	-6069	-5525
$\boldsymbol{\beta_1}$	100	0.1558	0.02802	0.1521	0.1595
$\boldsymbol{\beta_2}$	100	-15.8	3.612	-16.3	-15.3
$\boldsymbol{\beta_3}$	100	530	151.2	510	550
$\boldsymbol{\beta_4}$	64.6	0.02613	0.0232	0.02303	0.02922
$\boldsymbol{\beta_5}$	87.2	-0.3088	0.1691	-0.3314	-0.2862
$\boldsymbol{\beta_6}$	65.9	69.41	58.9	61.56	77.26

The inclusion of 95% credible intervals and posterior probabilities enhances the interpretability of the model, offering a comprehensive view of both the strength and reliability of each predictor's impact. Eq. 14 represents the final BMA model obtained. It shows that in the final BMA model, predictor variables are negatively and positively related to the response variable. These negatively and positively related variables are consistent with those in the regularization regression model.

$$\hat{y} = -5797 + 0.16x_3^3 - 15.8x_3^2 + 530.03x_3 + 0.03x_2x_3^2 - 0.32x_3^2x_5 + 69.41x_2^2x_5 \qquad (14)$$

4.5. Testing and Evaluation of Models

Four models have been trained: the ridge regression model, LASSO regression, elastic-net regression, and the final BMA model. The BMA model combines six selected polynomial regression models, with each model weighed according to its performance. These four models were subsequently tested using a separate test dataset. The purpose of testing the models is twofold: first, to evaluate the models' performance when applied to data they have not encountered before, and second, to assess the accuracy of each model in predicting the number of hotspots. This phase is critical in determining how well the models generalize beyond the training data and their reliability in making accurate predictions in real-world scenarios.

The performance of the four models was evaluated using two model evaluation metrics: Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). RMSE indicates how close the model's predictions are to the actual values, whereas a lower RMSE indicates better predictive accuracy. On the other hand, R^2 measures the proportion of variance in the dependent variable that can be explained by the model, with higher values indicating better explanatory power. Performance measurement based on RMSE and R^2 was conducted on the training and test datasets. The best-performing model was selected based on the lowest RMSE and the highest R^2 values. The RMSE and R^2 values for all four models are summarized in **Table 7**.

Table 7.

RMSE and R-squared values on training and testing data for regularized regression and BMA models.

Model	RMSE		R^2 (%)	
	Training	*Testing*	*Training*	*Testing*
Ridge	680,82	764,59	82,56	84,46
LASSO	680,04	752,31	82,6	85,06
Elastic-net	684,77	802,66	82,36	82,58
BMA	681,97	664,33	82,5	88,58

The values presented in **Table 7** indicate no significant differences in RMSE and R^2 values across the models when evaluated on the training data. However, these differences become more apparent when assessing the models on the test data. The final BMA model achieved the lowest RMSE on the test data, although this was not the case for the training data. Notably, the difference between the RMSE values for the BMA model on the training and test datasets was the smallest among all the models, suggesting that the BMA model is less likely to experience overfitting based on RMSE

values. Furthermore, the BMA model also had the highest R^2 value on the test data, reaching 88.58% and its R^2 value on the training data was also relatively high, at 82.5%. These R^2 values demonstrate that the BMA model can explain approximately 88.58% of the variance in the test data and 82.5% of the variance in the training data. Therefore, the BMA model outperforms the other three models and is the best predictor of the number of hotspots in Kalimantan.

BMA demonstrates superior robustness and accuracy in predictive modeling, particularly on test datasets, compared to regularization techniques such as Ridge, LASSO, and Elastic Net. Although these regularization methods effectively prevent overfitting and improve model interpretability by introducing penalties for complexity, they rely on a single model selection approach. This means that their predictive performance is contingent on the specific model chosen, which may not capture the full uncertainty inherent in the data. In contrast, BMA explicitly accounts for model uncertainty by averaging over a set of candidate models, weighted according to their posterior probabilities. This ensemble approach integrates information from multiple models, allowing BMA to leverage diverse perspectives on the data. As a result, it tends to produce more stable predictions, especially in scenarios where the underlying data-generating process is complex or not fully understood.

BMA incorporates prior information and allows for the inclusion of prior beliefs about model parameters, leading to more informed predictions. This characteristic enhances its flexibility and adaptability to various data distributions, improving its performance on unseen data. When regularization methods may inadvertently select suboptimal models, BMA mitigates this risk by pooling information across multiple models, thereby capturing a broader range of possible outcomes. Furthermore, BMA's reliance on a probabilistic framework enables it to provide measures of uncertainty alongside predictions, which is invaluable for decision-making in fields like environmental science and resource management. By quantifying uncertainty, practitioners can make more informed decisions based on the predicted hotspots for forest fires or other critical events.

While Ridge, LASSO, and Elastic Net may perform comparably during training, their single-model focus can limit their generalizability. In contrast, BMA's ensemble approach, which averages predictions across multiple models and incorporates uncertainty, enhances its robustness and predictive accuracy on test datasets.

4.6. Results Validation and Simulation

To ensure the reliability and robustness of the BMA model, a comprehensive validation process was conducted using climate data from 2021 to 2024. This step was critical to assess the model's performance on unseen data, ensuring its predictive accuracy and generalizability under varying climatic conditions. The validation involved applying the trained BMA model to a new dataset, which was processed consistently using the same methodology as the original dataset. Maintaining consistency in data preparation ensured that any observed differences in performance could be attributed solely to the model's capabilities, rather than inconsistencies in data handling.

The validation focused on key climate indicators that significantly influence fire risks. One such indicator was precipitation anomalies (x_2), which measure deviations in rainfall from the long-term average. This variable highlight unusual climatic conditions, such as abnormally dry or wet periods, that can impact fire susceptibility. Another crucial indicator was dry spells (x_3), defined as the number of days with less than 1 mm of rainfall over a specific period. Prolonged dry spells are strongly associated with heightened fire risk due to reduced soil moisture and increased vegetation flammability.

In addition to these local climate variables, the validation also incorporated the Indian Ocean Dipole (IOD) index (x_5), a large-scale climate indicator that measures differences in sea surface temperatures between the western and eastern Indian Ocean. The IOD index has a well-documented influence on regional weather patterns, including rainfall variability across Kalimantan. By including this index, the BMA model was able to account for broader climatic drivers that impact local fire dynamics. Pre-processed data, summarized in **Figure 5**, provides monthly data for three key climate indicators over the period from January 2021 to September 2024: precipitation anomalies (x_2), dry spells (x_3), and the Indian Ocean Dipole index (x_5).

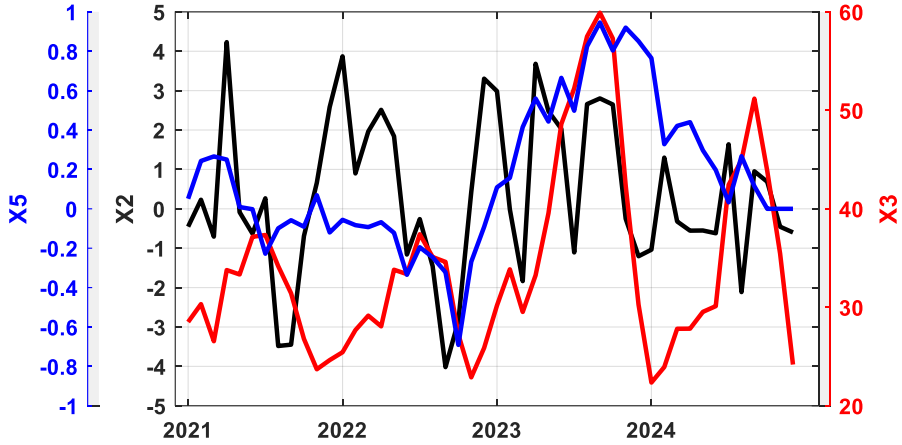


Fig. 5. Pre-processed data for results validation, covering precipitation anomalies (x_2), dry spells (x_3), and the Indian Ocean Dipole index (x_5) from 2021-2024.

Notable dry periods with prolonged dry spells and negative precipitation anomalies occurred in mid-2023 to end-2023, coinciding with positive IOD phases that exacerbate dry conditions. Conversely, wetter conditions with positive precipitation anomalies and fewer dry days were observed during 2022 and parts of 2024. The highest dry spell counts in mid-2023 and increasing dryness toward 2024 highlight periods of elevated fire susceptibility. Using the trained BMA model, validation data from 2021 to 2024 was employed to predict the occurrence of forest fire hotspots in Kalimantan and the results can be seen in **Figure 6**.

Figure 6 displays the predicted number of forest fire hotspots in 2021 to 2024, modeled using BMA. The black line indicates observed historical hotspot data, with a notable spike in late 2019 representing an extreme fire event. This historical pattern was likely used to calibrate its predictive capabilities. Expected values (red line) represents the central predictions of the model for the number of hotspots. It shows a generally low level of fire activity from 2021 to early 2023, with a clear upward trend and peak in mid-to-late 2023. Uncertainty range (yellow shading) represents the 95% confidence interval (CI). This highlights the uncertainty in the model's predictions, widening significantly during periods of high hotspot activity, such as the peaks in 2023 and 2024.

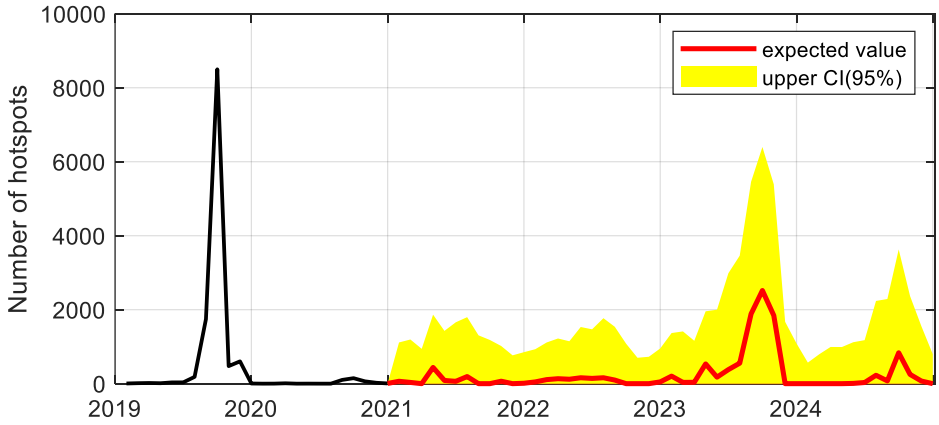


Fig. 6. Predicted number of forest fire hotspots in 2021 to 2024, modeled using BMA.

Predictions suggest a stable and relatively quiet period in terms of hotspots during 2021 and 2022. Starting in mid-2023, the model predicts an increase in fire activity, peaking in late 2023. These peaks

are likely tied to climatic conditions, such as prolonged dry spells or anomalies in rainfall, which are key predictors in the BMA model. During periods of low predicted activity (e.g., 2021-2022), the confidence intervals are narrow, reflecting high model confidence. In contrast, during peak activity (e.g., 2023-2024), the uncertainty widens, indicating the model's acknowledgment of higher variability and unpredictability under extreme conditions.

We compared our results with burned area data from SiPongi KLHK (see https://sipongi.menlhk.go.id/). **Table 8** shows burned area data from SiPongi highlights trends in fire activity across Kalimantan provinces between 2019 and 2023. In 2019, the burned area reached an exceptional total of 684,599 hectares, driven by extensive fires across Kalimantan, especially in Kalimantan Tengah (317,749 hectares). This aligns with the observed spike in hotspots in the historical data used for testing the BMA model. In 2020, the burned area dropped sharply to 26,286 hectares, reflecting a significant reduction in fire activity. This decrease is also reflected in the dramatic reduction in observed hotspots. These trends validate the model's ability to respond to extreme events and to capture transitions between high and low fire activity periods.

Table 8.

Burned area data from SiPongi highlights trends in fire activity across Kalimantan provinces between 2019 and 2023.

Region	2019	2020	2021	2022	2023
Kalimantan Barat	151,919	7,646	20,590	21,836	111,848.4
Kalimantan Selatan	137,848	4,017	8,625	429	190,394.6
Kalimantan Tengah	317,749	7,681	3,653	1,554	165,896.4
Kalimantan Timur	68,524	5,221	3,029	373	39,494.4
Kalimantan Utara	8,559	1,721	1,678	370	796.4
Total (in hectares)	**684,599**	**26,286**	**37,575**	**24,562**	**508,430.2**

During 2021 and 2022, the burned area remained relatively low, at 37,575 hectares and 24,562 hectares, respectively. The BMA model's predictions for this period indicate relatively low numbers of hotspots with narrow confidence intervals, reflecting strong confidence in its forecasts. This alignment with SiPongi data demonstrates that the model effectively captures periods of reduced fire activity, particularly under stable climatic conditions and low fire risk.

In 2023, a sharp increase in burned area is recorded, totaling 508,430 hectares. The most affected provinces were Kalimantan Barat (111,848 hectares), Kalimantan Selatan (190,394 hectares), and Kalimantan Tengah (165,896 hectares). Many recent studies have mentioned this fire incident in their studies (Nurlatifah et al., 2025). Correspondingly, the BMA model predicts a significant rise in hotspots for the same period, with notable peaks in expected values and wider confidence intervals. This reflects the model's awareness of heightened fire risks, and the variability associated with extreme fire seasons. The alignment between the burned area data and hotspot predictions in 2023 validates the model's utility for forecasting high-risk periods.

Furthermore, we simulate the predictions of hotspots using the trained BMA model under neutral and positive Indian Ocean Dipole (IOD) conditions to reveal the relationship between predictor variables x_2 (negative rainfall anomaly) and x_3 (number of dry days) with the predicted number of hotspots. **Figure 7** shows simulation results for various values of x_2 and x_3 under neutral Indian Ocean Dipole (IOD) conditions to predict hotspots in Kalimantan. Meanwhile, **Figure 8** displays simulation results under positive IOD.

Figure 7a illustrates how the number of hotspots increases as x_3 (number of dry days) rises from 25 to 60, under different fixed values of x_2 ranging from -5 to 5. A consistent trend is observed where higher values of x_3 led to a rapid increase in predicted hotspots, particularly when x_3 exceeds 50. This indicates that prolonged periods of dryness have a significant amplifying effect on fire activity. Moreover, for any given x_3, higher values of x_2 (indicating lower rainfall anomalies or drier conditions) result in more hotspots. The steep increase in hotspots with both variables highlights the compounded impact of extended dry periods and reduced rainfall on fire risk.

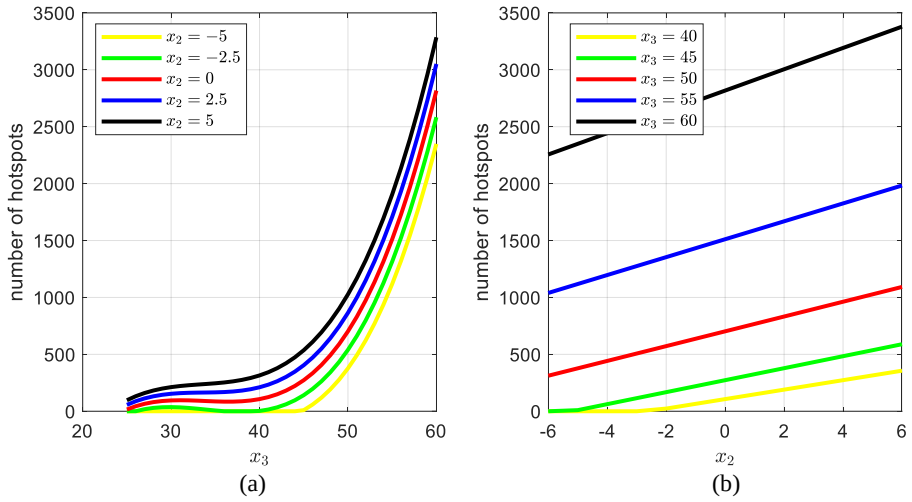


Fig. 7. Simulation results for various values of x_2 and x_3 under neutral Indian Ocean Dipole (IOD) conditions to predict hotspots in Kalimantan: a) influence of x_3 on the number of hotspots under different fixed values of x_2, and b) influence of x_2 on the number of hotspots under different fixed values of x_3.

Otherwise, **Figure 7b** examines the influence of x_2 on the number of hotspots while keeping x_3 constant at values ranging from 40 to 60. For all x_3 scenarios, the number of hotspots increases as x_2 moves from negative (wet conditions) to positive (dry conditions). Notably, the rate of increase in hotspots is steeper for higher x_3 values. For instance, at $x_3 = 60$, even a small shift in x_2 towards drier conditions leads to a marked increase in hotspots, emphasizing the vulnerability of regions with extended dry days to changes in rainfall anomalies.

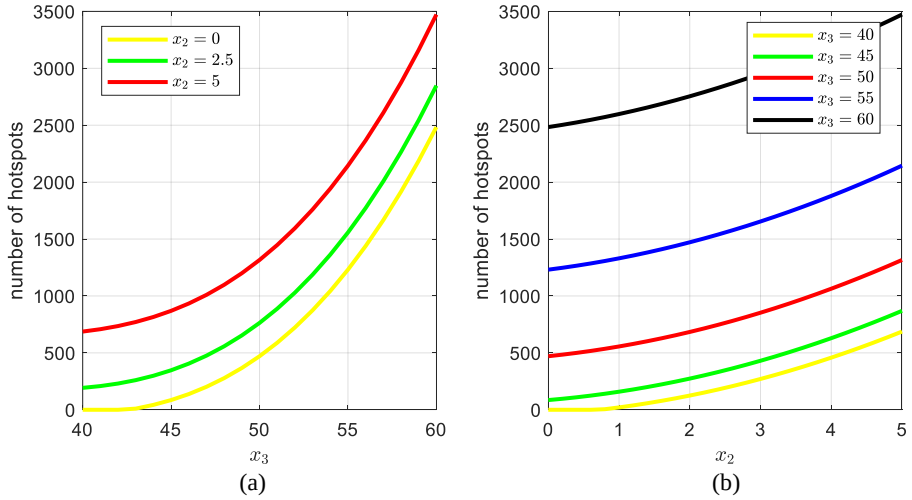


Fig. 8. Simulation results for various values of x_2 and x_3 under positive Indian Ocean Dipole (IOD) conditions to predict hotspots in Kalimantan: a) influence of x_3 on the number of hotspots under different fixed values of x_2, and b) influence of x_2 on the number of hotspots under different fixed values of x_3.

Under positive IOD, the value of x_2 and x_3 generally above their normal conditions. Therefore, the simulated range of x_2 and x_3 will be limited to values above their normal.

Figure 8a shows that as x_3 increases from 40 to 60, the predicted number of hotspots rises steadily for all fixed values of x_2. This indicates that under positive IOD conditions, a longer duration of dry days significantly escalates fire activity. The increase is more pronounced for higher values of x_2 (representing lower rainfall anomalies). For instance, when $x_2 = 5$, the growth in hotspots with increasing x_3 is considerably steeper than for lower x_2 values, highlighting the compounded impact of dry conditions and extended periods without rain. Otherwise, **Figure 8b** examines the impact of x_2 on the number of hotspots for fixed x_3 values ranging from 40 to 60. The number of hotspots increases as x_2 becomes more positive, with the rate of increase being greater for higher x_3 values. For example, at $x_3 = 60$, even small increments in x_2 lead to a substantial rise in predicted hotspots. This result indicates that under positive IOD conditions, the combination of a higher number of dry days and reduced rainfall intensifies fire risk more than each variable individually.

The simulations under both neutral and positive Indian Ocean Dipole (IOD) conditions provide valuable insights into the relationships between the predictor variables x_2 (negative rainfall anomaly) and x_3 (number of dry days) and the predicted number of hotspots. These results highlight distinct patterns influenced by different climatic conditions, which are crucial for understanding and mitigating fire risks in Kalimantan. In **both neutral and positive IOD** conditions, the number of hotspots consistently increases with higher x_3, reflecting the significant role of prolonged dry spells in intensifying fire activity. Under neutral IOD conditions, this relationship is evident but less pronounced compared to positive IOD scenarios. Positive IOD conditions amplify the effect of dry days, with hotspots increasing more steeply as x_3 rises from 40 to 60. This suggests that during positive IOD phases, the combination of regional weather patterns and longer dry periods creates a more conducive environment for fires. Conversely, the simulations show that the number of hotspots increases as x_2 becomes more positive (indicating lower-than-normal rainfall), with the effect being stronger **under positive IOD** conditions. For fixed x_3 values, higher x_2 results in a more rapid escalation of hotspots, particularly when x_3 is already elevated. This compounding effect of reduced rainfall and extended dry periods underscores the critical importance of monitoring rainfall anomalies in predicting fire risk.

The interaction between x_2 and x_3 is particularly noteworthy. During neutral IOD conditions, both variables influence the number of hotspots, but their combined effect is less severe compared to positive IOD conditions. Under positive IOD scenarios, the simultaneous increase in x_2 and x_3 leads to a dramatic rise in predicted hotspots. For instance, when x_3 is at its highest values (e.g., 60), even small increments in x_2 can significantly amplify the number of hotspots. This highlights the synergistic impact of these two variables during positive IOD phases, making such conditions particularly dangerous for fire outbreaks.

These combined results underscore the heightened fire risk during positive IOD conditions due to the stronger influence of both dry spells and rainfall anomalies. While neutral IOD conditions also present fire risks, the amplified effect during positive IOD highlights the importance of tailored fire mitigation strategies based on prevailing climatic conditions. Monitoring both x_2 and x_3 is critical, especially during positive IOD phases, as their interaction can significantly elevate fire activity. Proactive measures, including early warnings and fire prevention efforts, should prioritize areas experiencing prolonged dry spells and below-normal rainfall during these periods.

4.7. Discussion and Limitations of the Study

The findings underscore the practical utility of the BMA model for forest fire management. By accurately predicting hotspots, this model empowers stakeholders to make proactive and data-driven decisions to mitigate forest fire risks. For instance, the integration of climate indicators, particularly dry spells and precipitation anomalies, enables the identification of critical periods when fire risks are heightened due to prolonged dry conditions or abnormal weather patterns. This temporal framework is crucial for optimizing the allocation of resources, such as positioning firefighting teams, enhancing surveillance in high-risk areas, and pre-positioning water storage for firefighting.

Furthermore, the model's predictive capabilities can be integrated into early warning systems, providing timely alerts to local communities and allowing for the implementation of preventive measures such as controlled burns or the temporary suspension of activities like land clearing.

The ability of the BMA model to incorporate uncertainty into its predictions enhances its robustness compared to single-model approaches. By providing probabilistic insights, the BMA model enables stakeholders to assess the likelihood of fire outbreaks under different scenarios. This is particularly important in a complex and dynamic environment like Kalimantan, where variability in climate conditions can lead to unforeseen challenges. By quantifying these uncertainties, decision-makers can implement risk-informed strategies, such as prioritizing areas with a higher probability of fire occurrence while maintaining preparedness for lower-risk zones. This approach minimizes the potential for over-preparedness, which could result in resource wastage, and under-preparedness, which might exacerbate fire impacts. Ultimately, the BMA model equips policymakers and environmental managers with a tool not only for accurate forecasting but also for devising flexible and adaptive fire management strategies that can be dynamically adjusted as conditions evolve.

While the study demonstrates the effectiveness of the BMA model in predicting forest fire hotspots based on temporal climate indicators, it does not incorporate spatial aspects into the prediction framework. This limitation arises because the analysis focuses on temporal patterns, such as dry spells, precipitation anomalies, and climatic indices, without accounting for the geographic variability of these factors across Kalimantan. For example, regions with distinct ecological and topographical characteristics may respond differently to the same climatic conditions, leading to spatial heterogeneity in hotspot occurrences.

The absence of spatial considerations in the model implies that hotspot predictions are generalized across the entire study area, potentially overlooking localized factors such as land cover type, vegetation density, and proximity to human activities. Incorporating spatial data in future studies, such as GIS-based mapping or geostatistical models, could provide a more comprehensive understanding of fire dynamics and improve the model's utility for targeted interventions. For instance, spatially explicit models could help prioritize specific areas for monitoring or intervention, based on both temporal and spatial risk factors.

5. CONCLUSION

The first objective of the study was to identify the best combination of predictors for forecasting forest fire hotspots in Kalimantan based on climate indicators. The selection process aimed at minimizing the Bayesian Information Criterion (BIC) value, ensuring that only the most relevant variables were used. The research successfully used the best subset selection method, incorporating three key variables: precipitation anomalies, dry spells, and the Indian Ocean Dipole (IOD) index, into a mathematical equation consisting of six terms. The results showed that dry spells were a critical factor in predicting the occurrence of fire hotspots, playing a role in almost every significant model term. The use of polynomial interactions among these predictors allowed for a more nuanced understanding of how climate variables interrelate in driving fire risks.

The second goal was to construct and compare multiple regularized regression models, including Ridge, LASSO, and Elastic Net, with a BMA. The regularized regression models aim to address common issues like multicollinearity and overfitting by introducing penalties for complexity. While Ridge, LASSO, and Elastic Net performed adequately, the study found that these single-model approaches had limitations, particularly in their ability to generalize well to new data. In contrast, the BMA model, which averages predictions from multiple models weighted by their posterior probabilities, offered a more comprehensive approach by integrating the benefits of various models and accounting for the inherent uncertainty in the data.

Lastly, the study aimed to determine the best-performing model by evaluating the predictive accuracy of each model using Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). Testing on unseen data revealed that the BMA model outperformed all other models, achieving

the lowest RMSE (664) and the highest R^2 (88.58%) on test data. The BMA model's ensemble approach proved more robust, producing more accurate and reliable predictions, especially in the complex, uncertain scenario of predicting hotspots. This makes BMA a valuable tool for forest fire prediction, particularly in regions like Kalimantan, where climate conditions vary significantly.

The study's results offer promising prospects for implementing more accurate predictive models in forest fire management in Kalimantan, particularly through the superior BMA approach over regularized regression models. Its ability to forecast hotspots with high accuracy provides a strategic advantage for early intervention, resource prioritization, and mitigation planning. For policymakers, the model's predictions can inform proactive measures, such as community evacuation plans, policy adjustments regarding land use during critical periods, and optimizing budget allocation for firefighting efforts. This enables policymakers to predict fire hotspots, paving the way more precisely for effective preventive measures and improved early warning systems.

This study highlights the strength of using BMA for predicting forest fire hotspots based on temporal climate indicators. However, it is important to acknowledge that the model's focus is limited to temporal dynamics, without considering spatial variability in fire occurrence across Kalimantan. This limitation restricts the model's applicability for geographically targeted interventions, which are crucial for efficient resource allocation and localized risk management. Future research should aim to integrate spatial data, such as land use patterns, vegetation types, and geographic features, to enhance the predictive power and applicability of the model. Incorporating these spatial dimensions would enable the development of a more holistic fire risk assessment framework, combining both temporal and spatial predictors to better inform decision-making processes in forest fire management

Future studies could enhance this application by integrating real-time data streams, enabling dynamic updates to risk assessments and further improving the practicality of the model in rapidly changing environmental conditions. The BMA can also be extended to include a wider range of models and not just focus on polynomial models. These models can be tree-based, non-linear, machine learning, probabilistic models like Gaussian process, to neural networks. Moreover, further research should explore incorporating additional climate-related variables and utilizing advanced remote sensing technologies, as well as employing complex machine learning methods like deep learning. Developing adaptive predictive models that account for dynamic climate changes and new variables, such as land use effects, will be crucial for enhancing environmental management policies in Indonesia and other tropical regions.

ACKNOWLEDGEMENTS

We thank the Department of Mathematics, IPB University and the anonymous reviewers for their support throughout this research.

Data availability. Data generated or analysed during this study is provided within this manuscript or available at GitHub site: https://github.com/mkhoirun-najiboi/BMA-hotspots.

Code availability. The codes used during the current study are available at GitHub site: https://github.com/mkhoirun-najiboi/BMA-hotspots.

REFERENCES

Alhassan, E., Rochman, D., Schnabel, G., & Koning, A. J. (2024). Bayesian Model Averaging (BMA) for nuclear data evaluation. *Nuclear Science and Techniques*, 35(11), 1-26. https://doi.org/10.1007/s41365-024-01543-w

Alkhatib, R., Sahwan, W., Alkhatieb, A., & Schütt, B. (2023). A Brief Review of Machine Learning Algorithms in Forest Fires Science. *Applied Sciences (Switzerland)*, 13(14). https://doi.org/10.3390/app13148275

Andana, A. P., Safitri, D., & Rusgiyono, A. (2017). Model regresi menggunakan least absolute shrinkage and selection operator (lasso) pada data banyaknya gizi buruk kabupaten/kota di Jawa Tengah. *Jurnal Gaussian*, 6(1), 21–30.

Barros, A. M. G., Day, M. A., Preisler, H. K., Abatzoglou, J. T., Krawchuk, M. A., Houtman, R., & Ager, A. A. (2021). Contrasting the role of human- And lightning-caused wildfires on future fire regimes on a Central Oregon landscape. *Environmental Research Letters*, 16(6). https://doi.org/10.1088/1748-9326/ac03da

Basha, S. M., & Rajput, D. S. (2019). Survey on Evaluating the Performance of Machine Learning Algorithms: Past Contributions and Future Roadmap. In Deep Learning and Parallel Computing Environment for Bioengineering Systems (pp. 153–164). *Academic Press*. https://doi.org/10.1016/B978-0-12-816718-2.00016-6

Bertsimas, D., & Wiberg, H. (2020). Machine Learning in Oncology: Methods, Applications, and Challenges. *JCO Clinical Cancer Informatics*, 4, 885–894. https://doi.org/10.1200/cci.20.00072

Borrego, C., Miranda, A. I., Carvalho, A. C., & Carvalho, A. (2025). Forest fires and air pollution: A local and a global perspective. *WIT Transactions on Ecology and the Environment*, 37. https://doi.org/10.2495/AIR990711

Brasika, I. B. M., Antara, I. M. O. G., & Karang, I. W. G. A. (2021). Investigating El Nino Southern Oscillation as the main driver of forest fire in Kalimantan. *Malaysian Journal of Society and Space*, 17(4). https://doi.org/10.17576/geo-2021-1704-21

Brooks, G. P., & Ruengvirayudh, P. (2016). Best subset selection criteria for multiple linear regression. *General Linear Model Journal*, 42(2), 14–25.

Chi, H., Wu, Y., Zheng, H., Zhang, B., Sun, Z., Yan, J., ... & Guo, L. (2023). Spatial patterns of climate change and associated climate hazards in Northwest China. *Scientific Reports*, 13(1), 10418. https://doi.org/10.1038/s41598-023-37349-w

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, 1–24. https://doi.org/10.7717/PEERJ-CS.623

Claeskens, G., & Hjort, N. L. (2008). Frequentist and Bayesian model averaging. In *Model Selection and Model Averaging* (pp. 192–226). Cambridge University Press. https://doi.org/10.1017/CBO9780511790485.008

Dziak, J. J., Coffmann, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, 21(2), 553–565. https://doi.org/10.1093/bib/bbz016

Edwards, R. B., Naylor, R. L., Higgins, M. M., & Falcon, W. P. (2020). Causes of Indonesia's forest fires. *World Development*, 127. https://doi.org/10.1016/j.worlddev.2019.104717

Ertugrul, M., Varol, T., Ozel, H. B., Cetin, M., & Sevik, H. (2021). Influence of climatic factor of changes in forest fire danger and fire season length in Turkey. *Environmental monitoring and assessment*, 193, 1-17. https://doi.org/10.1007/s10661-020-08800-6

Fanin, T., & Van Der Werf, G. R. (2017). Precipitation-fire linkages in Indonesia (1997-2015). *Biogeosciences*, 14(18), 3995–4008. https://doi.org/10.5194/bg-14-3995-2017

Fikri, A. F., Agwil, W., & Agustina, D. (2023). Performa Teknik Regularisasi Dalam Penanganan Masalah Multikolinieritas. *Diophantine Journal of Mathematics and Its Applications*, 2(1), 45–51. https://doi.org/10.33369/diophantine.v2i01.28480

Gaveau, D. L. A., Sloan, S., Molidena, E., Yaen, H., Sheil, D., Abram, N. K., Ancrenaz, M., Nasi, R., Quinones, M., & Wielaard, N. (2014). Four decades of forest persistence, clearance and logging on Borneo. *PLoS ONE*, 9(7), 1–11. https://doi.org/10.1371/journal.pone.0101654

Giglio, L., Boschetti, L., Roy, D. P., Humber, M. L., & Justice, C. O. (2018). The Collection 6 MODIS burned area mapping algorithm and product. *Remote Sensing of Environment*, 217, 72–85. https://doi.org/10.1016/j.rse.2018.08.005

Han, J., Pei, J., & Tong, H. (2024). *Data Mining: Concepts and Techniques*, 4th ed. (Fourth Edi). Morgan Kaufmann. https://doi.org/10.1016/C2013-0-18660-6

Handayani, A., & Wachidah, L. (2022). Metode Regresi Elastic-Net untuk Mengatasi Masalah Multikolinearitas pada Kasus Tingkat Pengangguran Terbuka di Provinsi Jawa Barat. *Bandung Conference Series: Statistics*, 2(2), 459–465. https://doi.org/10.29313/bcss.v2i2.4722

Hardiyanti, O., & Nurmanina, A. (2020). Analysis of The Utilization of the Social Center for Orangutan Protection (COP) In Kalimantan in Orangutan Saving Efforts. *Progress In Social Development*, 1(1), 9–17. https://doi.org/10.30872/psd.v1i1.14

Harrison, M. E., Deere, N. J., Imron, M. A., Nasir, D., Adul, Asti, H. A., Soler, J. A., Boyd, N. C., Cheyne, S. M., Collins, S. A., D'Arcy, L. J., Erb, W. M., Green, H., Healy, W., Hendri, Holly, B., Houlihan, P. R., Husson, S. J., Iwan, … Struebig, M. J. (2024). Impacts of fire and prospects for recovery in a tropical peat forest ecosystem. *Proceedings of the National Academy of Sciences of the United States of America*, 121(17). https://doi.org/10.1073/pnas.2307216121

Hastie, T., Tibshirani, R., & Tibshirani, R. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science*, 35(4), 579–592. https://doi.org/10.1214/19-STS733

Herawati, N., Nisa, K., & Setiawan, E. (2018). Regularized Multiple Regression Methods to Deal with Severe Multicollinearity. *International Journal of Statistics and Applications*, 8(4), 167–172.

Hidayat, M. N., Wafdan, R., Iskandar, T., Dewi, C. D., Nurhayati, N., Ramli, M., ... & Rizal, S. (2025). The influence of El Niño-Southern Oscillation and the Indian Ocean Dipole on chlorophyll-a in the Aceh waters during 1999–2023. *Int. Journal of Remote Sensing*, 1-21. https://doi.org/10.1080/01431161.2025.2460244

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E. J. (2020). A Conceptual Introduction to Bayesian Model Averaging. *Advances in Methods and Practices in Psychological Science*, 3(2), 200–215. https://doi.org/10.1177/2515245919898657

Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58, 54–59.

Hope, T. M. H. (2020). Chapter 4 - Linear regression. In A. Mechelli & S. Vieira (Eds.), *Machine Learning* (pp. 67–81). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-815739-8.00004-3

Hu, Y., Fernandez-Anez, N., Smith, T. E. L., & Rein, G. (2018). Review of emissions from smouldering peat fires and their contribution to regional haze episodes. *International Journal of Wildland Fire*, 27(5), 293–312. https://doi.org/10.1071/WF17084

Huang, T., & Merwade, V. (2023). Uncertainty analysis and quantification in flood insurance rate maps using Bayesian model averaging and hierarchical BMA. Journal of Hydrologic Engineering, 28(2), 04022038. https://doi.org/10.1061/JHYEFF.HEENG-5851

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, 14(12). https://doi.org/10.1088/1748-9326/ab4e55

Iskandar, I., Lestari, D. O., Saputra, A. D., Setiawan, R. Y., Wirasatriya, A., Susanto, R. D., ... & Kunarso. (2022). Extreme positive Indian Ocean Dipole in 2019 and its impact on Indonesia. *Sustainability*, 14(22), 15155. https://doi.org/10.3390/su142215155

Jain, P., Coogan, S. C. P., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4), 478–505. https://doi.org/10.1139/er-2020-0019

Javeed, A., Dallora, A. L., Berglund, J. S., Ali, A., Ali, L., & Anderberg, P. (2023). Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. *Journal of Medical Systems*, 47(1). https://doi.org/10.1007/s10916-023-01906-7

Jdey, I., Hcini, G., & Ltifi, H. (2023). Deep Learning and Machine Learning for Malaria Detection: Overview, Challenges and Future Directions. *International Journal of Information Technology and Decision Making*. https://doi.org/10.1142/S0219622023300045

Jolly, C. J., Dickman, C. R., Doherty, T. S., van Eeden, L. M., Geary, W. L., Legge, S. M., ... & Nimmo, D. G. (2022). Animal mortality during fire. *Global Change Biology*, 28(6), 2053-2065. https://doi.org/10.1111/gcb.16044

Kadir, E. A., Kung, H. T., AlMansour, A. A., Irie, H., Rosa, S. L., & Fauzi, S. S. M. (2023). Wildfire Hotspots Forecasting and Mapping for Environmental Monitoring Based on the Long Short-Term Memory Networks Deep Learning Algorithm. *Environments - MDPI*, 10(7). https://doi.org/10.3390/environments10070124

Kasali, J., & Adeyemi, A. A. (2022). Model-Data Fit using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and The Sample-Size-Adjusted BIC. *Square : Journal of Mathematics and Mathematics Education*, 4(1), 43–51. https://doi.org/10.21580/square.2022.4.1.11297

Keong, C. Y., & Onuma, A. (2021). Transboundary ecological conservation, environmental value and environmental sustainability: Lessons from the heart of borneo. *Sustainability (Switzerland)*, 13(17). https://doi.org/10.3390/su13179727

Kumar, S., & Kumar, A. (2022). Hotspot and trend analysis of forest fires and its relation to climatic factors in the western Himalayas. *Natural Hazards*, 114(3), 3529–3544. https://doi.org/10.1007/s11069-022-05530-5

Kurniadi, A., Weller, E., Min, S. K., & Seong, M. G. (2021). Independent ENSO and IOD impacts on rainfall extremes over Indonesia. *International Journal of Climatology*, 41(6), 3640-3656. https://doi.org/10.1002/joc.7040

Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods. BMC Bioinformatics, 15(1). https://doi.org/10.1186/1471-2105-15-8

Latif, S. D., Alyaa Binti Hazrin, N., Hoon Koo, C., Lin Ng, J., Chaplot, B., Feng Huang, Y., El-Shafie, A., & Najah Ahmed, A. (2023). Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches. *Alexandria Engineering Journal*, 82, 16–25. https://doi.org/10.1016/j.aej.2023.09.060

Mahendra, A. P., Pradipta, D., Saputro, Moh. R. B., & Kusrini, K. (2022). Application of the Decision Tree Method to Forest Fire Detection (Case Study: in Palembang, South Sumatra). *JTECS : Jurnal Sistem Telekomunikasi Elektronika Sistem Kontrol Power Sistem Dan Komputer*, 2(1), 75. https://doi.org/10.32503/jtecs.v2i1.2196

Margono, B. A., Potapov, P. V, Turubanova, S., Stolle, F., & Hansen, M. C. (2014). Primary forest cover loss in indonesia over 2000-2012. *Nature Climate Change*, 4(8), 730–735. https://doi.org/10.1038/nclimate2277

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, 1(2), 140–147. https://doi.org/10.38094/jastt1457

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). 7. Polynomial Regression Models. In *Introduction to Linear Regression Analysis*, 6th Edition (pp. 230–267). John Wiley & Sons.

Najib, M. K., Nurdiati, S., & Sopaheluwakan, A. (2021). Quantifying the joint distribution of drought indicators in Borneo fire-prone area. *IOP Conference Series: Earth and Environmental Science*, 880(1), 012002. https://doi.org/10.1088/1755-1315/880/1/012002

Najib, M. K., Nurdiati, S., & Sopaheluwakan, A. (2022a). Copula-based joint distribution analysis of the ENSO effect on the drought indicators over Borneo fire-prone areas. *Modeling Earth Systems and Environment*, 8(2), 2817–2826. https://doi.org/10.1007/s40808-021-01267-5

Najib, M. K., Nurdiati, S., & Sopaheluwakan, A. (2022b). Multivariate fire risk models using copula regression in Kalimantan, Indonesia. *Natural Hazards*, 113(2), 1263–1283. https://doi.org/10.1007/s11069-022-05346-3

Najib, M. K., Nurdiati, S., & Sopaheluwakan, A. (2024). Prediction of hotspots pattern in Kalimantan using copula-based quantile regression and probabilistic model: a study of precipitation and dry spells across varied ENSO conditions. *Vietnam Journal of Earth Sciences*, 46(1), 12–33. https://doi.org/10.15625/2615-9783/19302

Nugrahani, E. H., Nurdiati, S., Bukhari, F., Najib, M. K., Sebastian, D. M., & Fallahi, P. A. N. (2024). Sensitivity and feature importance of climate factors for predicting fire hotspots using machine learning methods. *IAES International Journal of Artificial Intelligence*, 13(2), 2210–2223. https://doi.org/10.11591/ijai.v13.i2.pp2212-2225

Nurdiati, S., Bukhari, F., Julianto, M. T., Najib, M. K., & Nazria, N. (2021). Heterogeneous Correlation Map Between Estimated ENSO And IOD From ERA5 And Hotspot In Indonesia. *Jambura Geoscience Review*, 3(2), 65–72. https://doi.org/10.34312/jgeosrev.v3i2.10443

Nurdiati, S., Bukhari, F., Julianto, M. T., Sopaheluwakan, A., Aprilia, M., Fajar, I., Septiawan, P., & Najib, M. K. (2022a). The impact of El Niño southern oscillation and Indian Ocean Dipole on the burned area in Indonesia. *Terrestrial, Atmospheric and Oceanic Sciences*, 33(15). https://doi.org/10.1007/S44195-022-00016-0

Nurdiati, S., Sopaheluwakan, A., Septiawan, P., & Ardhana, M. R. (2022b). Joint Spatio-Temporal Analysis of Various Wildfire and Drought Indicators in Indonesia. *Atmosphere*, 13(10). https://doi.org/10.3390/atmos13101591

Nurlatifah, A., Kombara, P. Y., Pratama, A., Faristyawan, R., Rakhman, A. A., & Noviastuti, N. (2025). Utilisation of WRF-HYSPLIT modelling approach and GEMS to identify PM2.5 sources in Central Kalimantan – study case: 2023 forest fire. *Journal of Southern Hemisphere Earth Systems Science*, 75(1). https://doi.org/10.1071/ES24006

Palamba, P. (2024). Understanding the environmental impacts of peatland fires: optical density, gas emissions and airflow effects. *International Journal of Environmental Engineering*, 12(4), 330–343. https://doi.org/10.1504/ijee.2024.10063659

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. *Materials Today: Proceedings*, 80, 3782–3785. https://doi.org/10.1016/j.matpr.2021.07.382

Preeti, T., Kanakaraddi, S., Beelagi, A., Malagi, S., & Sudi, A. (2021). Forest Fire Prediction Using Machine Learning Techniques. *2021 International Conference on Intelligent Technologies, CONIT 2021*. https://doi.org/10.1109/CONIT51480.2021.9498448

Purwanto, A., & Sudargini, Y. (2021). Partial Least Squares Structural Squation Modeling (PLS-SEM) Analysis for Social and Management Research : A Literature Review. *Journal of Industrial Engineering & Management Research*, 2(4), 114–123.

Rachman, H. A., Setiawati, M. D., Hidayah, Z., Syah, A. F., Nandika, M. R., Lumban-Gaol, J., ... & Syamsudin, F. (2024). Dynamic of upwelling variability in southern Indonesia region revealed from satellite data: Role of ENSO and IOD. Journal of Sea Research, 202, 102543. https://doi.org/10.1016/j.seares.2024.102543

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Sa'adi, Z., Shiru, M. S., Shahid, S., & Ismail, T. (2020). Selection of general circulation models for the projections of spatio-temporal changes in temperature of Borneo Island based on CMIP5. *Theoretical and Applied Climatology*, 139(1–2), 351–371. https://doi.org/10.1007/s00704-019-02948-z

Saharjo, B. H., & Hasanah, U. (2023). Analisis Faktor Penyebab Terjadinya Kebakaran Hutan Dan Lahan Di Kabupaten Pulang Pisau, Kalimantan Tengah. *Journal of Tropical Silviculture*, 14(1), 25–29. https://doi.org/10.29244/j-siltrop.14.01.25-29

Saharjo, B. H., & Nasution, M. R. A. (2021). Hotspot Distribution Pattern as an Indicator of Forest and Land Fires in West Aceh District. *Journal of Tropical Silviculture*, 12(2), 60–66. https://doi.org/10.29244/j-siltrop.12.2.60-66

Sahu, S. K., Mokhade, A., & Bokde, N. D. (2023). An Overview of Machine Learning, Deep Learning, and Reinforcement Learning-Based Techniques in Quantitative Finance: Recent Progress and Challenges. *Applied Sciences (Switzerland)*, 13(3). https://doi.org/10.3390/app13031956

Saleh, A. K. Md. E., Arashi, M., & Kibria, B. M. G. (2019). *Theory of Ridge Regression Estimation with Application*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118644478

Sambodo, N. P., Pradhan, M., Sparrow, R., & van Doorslaer, E. (2024). When the smoke gets in your lungs: short-term effects of Indonesia's 2015 forest fires on health care use. *Environmental Health: A Global Access Science Source*, 23(1). https://doi.org/10.1186/s12940-024-01079-x

Sanjaya, F. I., & Heksaputra, D. (2020). Prediksi Rerata Harga Beras Tingkat Grosir Indonesia dengan Long Short Term Memory. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, 7(2), 163–174. https://doi.org/10.35957/jatisi.v7i2.388

Sarker IH. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Syaufina, L., & Puspitasari, N. (2015). Correlation of Weather Factors and Forest Fire Occurence in KPH Bogor, Perum Perhutani Unit III West Java and Banten. *Journal of Tropical Silviculture*, 06(1), 43–48.

Tacconi, L. (2016). Preventing fires and haze in Southeast Asia. *Nature Climate Change*, 6(7), 640–643. https://doi.org/10.1038/nclimate3008

Tishbirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), 58(1), 267–288.

Usup, A., & Hayasaka, H. (2023). Peatland Fire Weather Conditions in Central Kalimantan, Indonesia. *Fire*, 6(5). https://doi.org/10.3390/fire6050182

van der Werf, G. R., Randerson, J. T., Giglio, L., van Leeuwen, T. T., Chen, Y., Rogers, B. M., Mu, M., van Marle, M. J. E., Morton, D. C., Collatz, G. J., Yokelson, R. J., & Kasibhatla, P. S. (2017). Global fire

emissions estimates during 1997-2016. *Earth System Science Data*, 9(2), 697–720. https://doi.org/10.5194/essd-9-697-2017

Venkatesh, K. A., Mishra, D., & Manimozhi, T. (2023). Model selection and regularization. In T. Goswami & G. R. Sinha (Eds.), *Statistical Modeling in Machine Learning* (pp. 159–178). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-323-91776-6.24001-3

von Rintelen, K., Arida, E., & Häuser, C. (2017). A review of biodiversity-related issues and challenges in megadiverse Indonesia and other Southeast Asian countries. *Research Ideas and Outcomes*, 3. https://doi.org/10.3897/rio.3.e20860

Wasilaine, T. L., Talakua, M. W., & Lesnussa, Y. A. (2014). Model Regresi Ridge Untuk Mengatasi Model Regresi Linier Berganda Yang Mengandung Multikolinieritas. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, 8(1), 31–37. https://doi.org/10.30598/barekengvol8iss1pp31-37

Yanke, A., Zendrato, N. E., & Soleh, A. M. (2022). Handling Multicollinearity Problems in Indonesia's Economic Growth Regression Modeling Based on Endogenous Economic Growth Theory. *Indonesian Journal of Statistics and Its Applications*, 6(2), 228–244. https://doi.org/10.29244/ijsa.v6i2p214-230

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. *Series B: Statistical Methodology*, 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x