Journal of Computation Physics and Earth Science
ISSN: 2776-2521 (online)
Volume 5, Number 1, April 2025, Page 15 - 20
https://journal.physan.org/index.php/jocpes/index

Comparative Analysis of Linear Regression Models and XGBoost to Assess the Impact of ENSO on Rainfall in Ternate City in 2023

Firman Almaliky Gapi Amra [1], Anton Widodo[2], Muchamad Rizqy Nugraha[3]
[1,2,3]Undergraduate Program in Applied of Instrumentation Meteorology, Climatology Geophysics (STMKG)

Article Info

Article history:

Received March 1, 2025
Revised March 6, 2025
Accepted March 7, 2025

Keywords:

ENSO,
Rainfall,
Linear Regression,
XGBoost,
Machine Learning.

ABSTRACT

The purpose of this study is to evaluate how well two prediction models—linear regression and XGBoost—perform in assessing how ENSO (El Niño-Southern Oscillation) affects rainfall in Ternate City in 2023. The Meteorology, Climatology, and Geophysics Agency (BMKG) provided monthly rainfall data, while the Bureau of Meteorology (BOM) in Australia provided ENSO index data. Performance indicators such Pearson correlation analysis, the coefficient of determination (R-squared), and mean squared error (MSE) were used in the evaluation. According to the findings, the two models perform differently when it comes to capturing the pattern of the link between rainfall and ENSO; XGBoost is more adaptable but has a tendency to overfit on small amounts of data, whereas linear regression obtains a better R-squared value.

Corresponden Author:

Firman Almaliky Gapi Amra,
Undergraduate Program in Applied of Instrumentation Meteorology, Climatology Geophysics (STMKG)
Tangerang City, Banten, Indonesia
Email: firman.amra14@gmail.com

1. INTRODUCTION

Climate change and its associated global phenomena have increasingly become critical areas of scientific research, with the El Niño-Southern Oscillation (ENSO) emerging as a particularly significant driver of regional climate variability[1], [2]. ENSO, encompassing both El Niño and La Niña phases, profoundly impacts global weather patterns, including precipitation, temperature, and extreme weather events[3].In Indonesia, a region characterized by complex meteorological dynamics, ENSO exerts profound influences on precipitation patterns, often resulting in significant societal and environmental challenges[4], [5]. Ternate City, situated in a climatologically sensitive area, serves as an exemplary case study for understanding these intricate relationships, particularly given its vulnerability to changing rainfall patterns and their associated risks. Understanding the relationship between ENSO and rainfall is crucial for improving predictive capabilities, enabling better resource management and disaster preparedness[6].Traditional statistical methods like linear regression have long been employed to analyze such relationships due to their simplicity and interpretability[7], [8]. Linear regression models the relationship between independent and dependent variables under the assumption of linearity, making it a foundational tool for identifying trends and quantifying relationships. By leveraging the straightforward assumptions of linear regression, researchers can establish a baseline understanding of how ENSO anomalies influence rainfall patterns, including the direct proportional changes between the ENSO index and precipitation levels[6], [9].[10]

However, the complex, non-linear nature of climate interactions often requires more advanced analytical approaches. XGBoost (Extreme Gradient Boosting), an ensemble machine learning algorithm, has emerged as a powerful tool for analyzing and predicting environmental phenomena. Unlike linear regression, XGBoost can model non-linear relationships and capture intricate feature interactions, which are often present in climate datasets. By iteratively constructing decision trees, XGBoost minimizes prediction errors while handling large datasets, missing values, and feature complexities effectively[11]. These advantages make it a suitable complement to traditional statistical methods in understanding the nuanced impacts of ENSO on regional precipitation[12], [13].

Machine learning techniques, including XGBoost, have revolutionized climate science by providing robust tools for analyzing environmental datasets. These methodologies enable the identification of hidden patterns in large and complex data, offering insights that traditional approaches may overlook. When applied to the ENSO-rainfall relationship in Ternate City, XGBoost has the potential to uncover subtle interactions that influence precipitation variability. Such insights are critical for refining predictive models, which can aid in climate adaptation and mitigation strategies, particularly in regions like Indonesia where the impacts of climate variability are pronounced[14].

The inherent complexity of ENSO's impact on regional precipitation necessitates a multi-model approach to accurately quantify and understand these dynamics. By comparing linear regression and XGBoost, this study aims to provide a comprehensive evaluation of their predictive capabilities[15], [16].While linear regression offers clarity and interpretability, XGBoost introduces flexibility and precision in capturing complex data patterns. This research not only seeks to quantify the relationship between ENSO anomalies and rainfall in Ternate City but also to demonstrate the broader applicability of machine learning in advancing climate science. Such comparative analyses are increasingly important as the need for accurate climate predictions grows in response to the accelerating pace of global climate change[17], [18].

The study seeks to comprehensively investigate the relationship between ENSO anomalies and rainfall in Ternate City by comparing the performance of linear regression and XGBoost models. Through a rigorous analytical approach, the research aims to evaluate the predictive capabilities of both methodologies using key performance metrics like Mean Squared Error and R-squared, ultimately drawing substantive conclusions about the most effective model for understanding and predicting rainfall patterns in the region. By systematically comparing these two modeling techniques, the study intends to provide insights into the complex dynamics of ENSO's influence on local precipitation and contribute to more accurate environmental forecasting strategies.

2. RESEARCH METHOD

This research uses ENSO anomaly data from BOM and monthly rainfall data in Ternate City throughout 2023 from BMKG. The analysis is conducted using the linear regression method and the XGBoost algorithm.
The Data Sourches for Monthly rainfall data obtained from **BMKG** (Meteorology, Climatology, and Geophysics Agency, ENSO anomaly index data obtained from **BOM** (Bureau of Meteorology, Australia) and Time Period: Year 2023.The Variables are Independent variable, ENSO anomaly index (from BOM) and Dependent variable: Monthly rainfall (from BMKG). The integrated modeling approach combines the strengths of Linear Regression and XGBoost to analyze the relationship between ENSO anomalies and rainfall. Linear Regression provides a straightforward initial understanding of the linear correlation, offering clear insights into direct relationships between variables. Meanwhile, XGBoost enhances the analysis by capturing complex non-linear patterns and interactions that traditional linear methods might overlook[19], [20]. The model performance in this study is comprehensively assessed using three key metrics: **Mean Squared Error (MSE)**, **R-squared (R^2)**, and **Pearson Correlation**. Together, these metrics provide a well-rounded evaluation of the models' ability to predict rainfall based on ENSO anomalies

Mean Squared Error (MSE) is a widely used metric for evaluating the accuracy of continuous prediction models. It measures the average squared difference between predicted values and actual observed values, with larger errors being penalized more heavily due to the squaring of differences. A lower MSE indicates better predictive accuracy, as it means that the model's predictions are closer to the true values. In this study, the linear regression model achieves a significantly lower MSE compared to the XGBoost model. This suggests that linear regression provides more accurate and reliable predictions for rainfall in Ternate City, based on ENSO anomalies. The higher MSE in the XGBoost model indicates that it struggles to make accurate predictions in this case, potentially due to overfitting to the limited dataset.

R-squared (R^2) is another critical evaluation metric, as it represents the proportion of variance in the dependent variable (rainfall) that can be explained by the independent variable (ENSO anomaly) within the model. R^2 values range from 0 to 1, where a value closer to 1 indicates a better fit of the model to the data. A higher R^2 suggests that the model explains a significant portion of the variability in the outcome, whereas a low R^2 indicates a poor model fit. In this study, the linear regression model shows an R^2 of 0.303, meaning it explains approximately 30.3% of the variation in rainfall based on the ENSO anomaly. While this is a modest value, it still demonstrates that the linear regression model is able to capture a meaningful portion of the relationship between ENSO anomalies and rainfall. In contrast, the R^2 value for the XGBoost model is negative (-0.059), indicating that the model fails to explain the variation in the data and actually performs worse than a simple baseline model that just predicts the mean of the rainfall values.

The Pearson correlation is used to assess the strength and direction of the linear relationship between predicted and actual values. A correlation coefficient closer to +1 indicates a strong positive linear relationship, while a value near

-1 suggests a strong negative relationship. A value near 0 indicates little to no linear relationship. In the context of this study, the Pearson correlation for the XGBoost model is -0.550, reflecting a weak negative correlation between the model's predictions and the actual rainfall data. This weak negative correlation suggests that XGBoost does not capture the relationship between ENSO anomalies and rainfall effectively. Although Pearson correlation for the linear regression model is not explicitly reported, it can be inferred that it is likely higher due to the model's better fit to the data and more accurate predictions.

Overall, these three metrics—MSE, R², and Pearson correlation—offer a multi-faceted evaluation of model performance. MSE provides insight into the accuracy of the predictions, R² indicates how well the model explains the variance in rainfall, and Pearson correlation evaluates the strength of the linear relationship between predicted and actual values. When taken together, these metrics clearly show that linear regression outperforms XGBoost in predicting rainfall based on ENSO anomalies. The results highlight the importance of selecting the appropriate model based on data characteristics, as linear regression demonstrates better generalization and accuracy with the available dataset.

3. RESULT AND DISCUSSION

In this section, we will systematically analyze and interpret the results of our comparative study on ENSO-rainfall relationships in Ternate City. Our investigation employs two distinct modeling approaches - linear regression and XGBoost - to unravel the complex dynamics of climate interactions.

Linear Regression Analysis Result

- **MSE**: 38.2921955135183
- **R-squared (R²)**: 0.3031723420952398
- **Regression Coefficient**: -4.417146709384934
- **Intercept**: 9.867402028082324

The linear regression results indicate that the relationship between ENSO anomaly and rainfall is negative. The regression line tends to decline, indicating that higher ENSO anomaly values correspond to decreased rainfall.

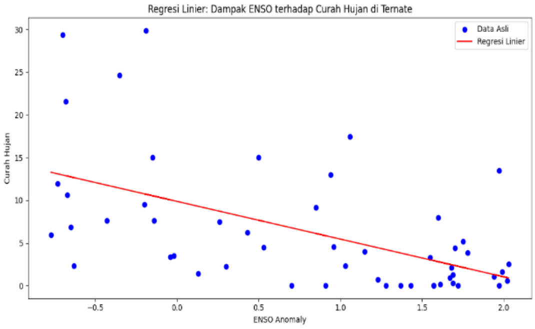

Figure 1. The linear regression results

The linear regression graph shows a negative trend between the ENSO anomaly index and rainfall. Each unit increase in the ENSO anomaly index tends to be followed by a decrease in rainfall, consistent with the negative regression coefficient obtained. This indicates that the El Niño phenomenon, characterized by positive ENSO indices, is associated with reduced rainfall in Ternate City.

XGBoost Analysis Results

- MSE: 88.3720951693223
- R-squared (R²): -0.05910609209615769
- Pearson Correlation: -0.5506108808362217

Although XGBoost offers higher flexibility, the results show that this model has a larger MSE compared to linear regression. The negative R-squared value indicates that the XGBoost model fails to explain the data variation effectively.

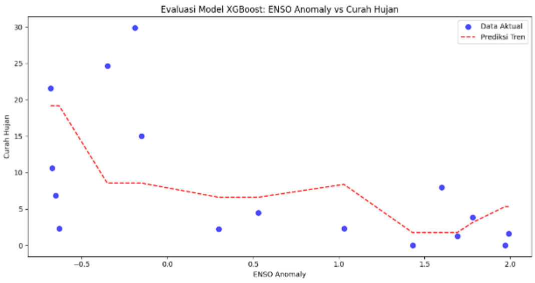

Figure 2. XGBoost Analysis Results

The XGBoost analysis graph shows more varied predictions but lacks consistency with the actual data pattern. The XGBoost model tends to overfit to the noise in the data, leading to higher MSE and a negative R-squared value. This suggests that under conditions with limited datasets like this, XGBoost fails to optimally capture the relationship between ENSO anomaly and rainfall.

TABLE I. MODEL COMPARISON

Evaluation Metric	*Linear Regression*	*XGBoost*
Mean Squared Error (MSE)	38.292	88,372
R-squared (R²)	0,303	-0.059
Pearson Correlation		-0.550

From Table I, it is evident that the **Mean Squared Error (MSE)** for the linear regression model is 38.292, which is significantly lower than the MSE of 88.372 for the XGBoost model. A smaller MSE indicates that the predictions from the linear regression model are closer to the actual data compared to the XGBoost model, which exhibits higher prediction errors.

The **R-squared (R²)** value for linear regression is 0.303, suggesting that 30.3% of the variance in rainfall can be explained by the linear relationship with the ENSO index. In contrast, XGBoost has a negative R² value (-0.059), indicating that the model fails to explain the variation in the data effectively. The negative R² also suggests potential overfitting or that the model does not fit the dataset well. Additionally, the **Pearson correlation coefficient** for XGBoost is -0.550, implying a weak negative correlation between the model's predictions and the actual data. This further supports the conclusion that XGBoost struggles to capture the relationship between ENSO anomalies and rainfall in this dataset. Overall, these results highlight that while XGBoost is generally effective in capturing complex patterns, its performance diminishes with limited datasets. In contrast, linear regression demonstrates better accuracy and generalizability in modeling the relationship between ENSO anomalies and rainfall.

4. CONCLUSION

This study highlights that Linear Regression is more effective than XGBoost in modeling the relationship between ENSO anomalies and rainfall in Ternate City. The superior performance of Linear Regression is evident from its significantly lower Mean Squared Error (MSE), indicating more accurate predictions, and its better generalization capabilities compared to XGBoost. While XGBoost is known for its ability to model complex relationships and capture intricate patterns in data, the limited dataset size in this study led to overfitting, which ultimately reduced its predictive reliability. The R-squared and Pearson correlation metrics further emphasized the strength of linear regression, as it demonstrated a meaningful relationship between ENSO anomalies and rainfall, explaining around

30.3% of the variation in rainfall. On the other hand, the negative R² and weak Pearson correlation for XGBoost suggested that it failed to effectively model the relationship in this contex.

The analysis also confirmed a negative correlation between ENSO anomalies and rainfall, as reflected by the negative regression coefficient in the linear regression model. This suggests that higher ENSO anomalies are associated with lower rainfall levels, supporting the established understanding that El Niño (characterized by positive ENSO anomalies) typically leads to reduced rainfall in many parts of the world, including Indonesia. These findings emphasize the importance of selecting the appropriate modeling approach based on the characteristics of the data. For datasets with limited observations and relatively simple relationships, classical statistical approaches like linear regression offer a more robust, interpretable, and reliable solution compared to more complex machine learning methods such as XGBoost. This study reinforces the value of linear regression in scenarios where data limitations and the simplicity of the relationship between variables make it a more suitable choice over advanced machine learning models.

REFERENCE

[1] H. S. Lee, “General Rainfall Patterns in Indonesia and the Potential Impacts of Local Seas on Rainfall Intensity,” *Water (Switzerland)*, vol. 7, no. 4, pp. 1751–1768, Apr. 2015, doi: 10.3390/w7041751.

[2] L. Agustina, R. H. Virgianto, and A. N. Fitrianto, “Utilization of remote sensing data for mapping the effect of Indian Ocean Dipole (IOD) and El Nino Southern Oscillation (ENSO) in Sumatra Island,” in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Nov. 2021. doi: 10.1088/1755-1315/893/1/012062.

[3] J. T. Nugroho, D. Nurfitriani, Suwarsono, G. A. Chulafak, R. J. Manalu, and S. Harini, “Rainfall anomalies assessment during drought episodes of 2015 in Indonesia using CHIRPS Data,” in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Apr. 2021. doi: 10.1088/1755-1315/739/1/012044.

[4] S. Nurdiati, F. Bukhari, A. Sopaheluwakan, P. Septiawan, and V. Hutapea, “ENSO AND IOD IMPACT ANALYSIS OF EXTREME CLIMATE CONDITION IN PAPUA, INDONESIA,” *Geographia Technica*, vol. 19, no. 1, pp. 1–18, Mar. 2024, doi: 10.21163/GT_2024.191.01.

[5] N. A. I. Baharuddin, M. Zainuddin, and Najamuddin, “The impact of ENSO-IOD on Decapterus spp. in Pangkajene Kepulauan and Barru Waters, Makassar Strait, Indonesia,” *Biodiversitas*, vol. 23, no. 11, pp. 5613–5622, 2022, doi: 10.13057/biodiv/d231110.

[6] A. Parmar, K. Mistree, and M. Sompura, “Machine Learning Techniques For Rainfall Prediction: A Review,” 2017. [Online]. Available: https://www.researchgate.net/publication/319503839

[7] D. Maulud and A. M. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning,” *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.

[8] “Simple Linear Regression Multiple Linear Regression Polynomial Regression Underfitting and Overfitting Implementing Linear Regression in Python Python Packages for Linear Regression Simple Linear Regression With scikit-learn Multiple Linear Regression With scikit-learn Polynomial Regression With scikit-learn Advanced Linear Regression With statsmodels Beyond Linear Regression Conclusion.” [Online]. Available: https://realpython.com/linear-regression-in-python/1/21

[9] T. Soares Dos Santos, D. Mendes, and R. Rodrigues Torres, “Artificial neural networks and multiple linear regression model using principal components to estimate rainfall over South America,” *Nonlinear Process Geophys*, vol. 23, no. 1, pp. 13–20, Jan. 2016, doi: 10.5194/npg-23-13-2016.

[10] I. Papailiou, F. Spyropoulos, I. Trichakis, and G. P. Karatzas, “Artificial Neural Networks and Multiple Linear Regression for Filling in Missing Daily Rainfall Data,” *Water (Switzerland)*, vol. 14, no. 18, Sep. 2022, doi: 10.3390/w14182892.

[11] N. Uzir, S. Raman, S. Banerjee, and R. S. Nishant Uzir Sunil R, “Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets Experimenting XGBoost Algorithm for Prediction and Classifi cation of Different Datasets,” *International Journal of Control Theory and Applications*, vol. 9, 2016, [Online]. Available: https://www.researchgate.net/publication/318132203

[12] A. Yasper, D. Handoko, M. Putra, H. K. Aliwarga, and M. S. R. Rosid, “Hyperparameters Optimization in XGBoost Model for Rainfall Estimation: A Case Study in Pontianak City,” *Jurnal Penelitian Pendidikan IPA*, vol. 9, no. 9, pp. 7113–7121, Sep. 2023, doi: 10.29303/jppipa.v9i9.3890.

[13] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[14] L. Ferreira, A. Pilastri, C. M. Martins, P. M. Pires, and P. Cortez, “A Comparison of AutoML Tools for Machine Learning, Deep Learning and XGBoost,” in *Proceedings of the International Joint Conference on*

Neural Networks, Institute of Electrical and Electronics Engineers Inc., Jul. 2021. doi: 10.1109/IJCNN52387.2021.9534091.

[15] Z. Kuang *et al.*, “A Hybrid ENSO Prediction System Based on the FIO−CPS and XGBoost Algorithm,” *Remote Sens (Basel)*, vol. 15, no. 7, Apr. 2023, doi: 10.3390/rs15071728.

[16] A. Kurniadi, E. Weller, S. K. Min, and M. G. Seong, “Independent ENSO and IOD impacts on rainfall extremes over Indonesia,” *International Journal of Climatology*, vol. 41, no. 6, pp. 3640–3656, May 2021, doi: 10.1002/joc.7040.

[17] A. Santoso, M. J. Mcphaden, and W. Cai, “The Defining Characteristics of ENSO Extremes and the Strong 2015/2016 El Niño,” Dec. 01, 2017, *Blackwell Publishing Ltd*. doi: 10.1002/2017RG000560.

[18] C. Wang, C. Deser, J. Y. Yu, P. DiNezio, and A. Clement, “El Niño and Southern Oscillation (ENSO): A Review,” in *Coral Reefs of the World*, vol. 8, Springer Nature, 2017, pp. 85–106. doi: 10.1007/978-94-017-7499-4_4.

[19] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, “Predicting motor insurance claims using telematics data—XGboost versus logistic regression,” *Risks*, vol. 7, no. 2, Jun. 2019, doi: 10.3390/risks7020070.

[20] J. Wu *et al.*, “Prediction and Screening Model for Products Based on Fusion Regression and XGBoost Classification,” *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/4987639.