Improving Rice Yield Prediction Accuracy using Regression Models with Climate Data

Mohamad Farhan Mohamad Mohsin[1][0000-0002-4393-7288], Muhammad Khalifa Umana[2][0000-0001-6678-6139], Mohamad Ghozali Hassan[3][0000-0001-9374-5711], Kamal Imran Mohd Sharif[4][0000-0003-0628-8319], Mohd Azril Ismail[5][0000-0003-0990-6653], [6]Khazainani Salleh [6][0009-0004-7087-8772], Suhaili Mohd Zahari[7][0009-0004-0822-4575], Mimi Adilla Sarmani[8][0009-0000-8243-8898], Neil Gordon [9][0000-0001-6889-0781]

[1,2] School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
[3,4,5] School of Technology Management and Logistics, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
[6 ,7,8] National Climate Centre, Malaysian Meteorological Department, 46667 Petaling Jaya, Selangor Darul Ehsan, Malaysia
[9]Computer Science, University of Hull, Hull, UK

[1]farhan@uum.edu.my, [2]umana_m_khalifa@ahsgs.uum.edu.my, [3]ghozali@uum.edu.my, [4]kamalimran@uum.edu.my, [5]azril@uum.edu.my, [6]khznani@met.gov.my,[7]suhaili@met.gov.my,[8]mimi_adilla@met.gov.my,[9] n.a.gordon@hull.ac.uk

Abstract.
Rice production is critical to food security, and accurate yield predictions are required for planning and decision-making. However, precisely predicting rice yields using machine learning models can be difficult due to the complicated interactions of various factors, such as how climate affects rice production. This study sought to solve this rice production is critical to food security, and accurate yield predictions are required for planning and decision-making. However, accurately predicting rice yields using machine learning models can be difficult due to the complicated interactions of various factors, such as how climate affects rice production. This study aims to address this issue by investigating how climate data affect Malaysian rice yield prediction models. The study used a linear regression model trained on rice production data and compared its performance with models incorporating climate data. Both datasets covered the period from 2010 to 2021 in Malaysia. The study found that including climate data significantly improved the prediction accuracy, with an approximately 77% improvement in MAE and 69% in RMSE. The results suggest that incorporating climate data into yield prediction models is essential for accurate and reliable predictions. These findings have important implications for stakeholders in the agricultural industry who can use accurate yield predictions to make informed

decisions. However, the study's limitations include using a single predictive model and data from a single country, suggesting the need for future studies to explore other machine learning algorithms and expand the scope of the research to other regions. Overall, this study contributes to the growing body of literature on the impact of climate data on yield prediction models and highlights the importance of considering climate data in agricultural decision-making.

Keywords: Rice production, Climate data, Machine learning, Crop yield prediction, Linear regression

1 Introduction

Rice is a staple food for many people worldwide, including Malaysians. Meeting domestic rice demand is difficult because of distracting variables such as rising population, changes in land use, soil quality, weather patterns, plantation diseases, and restricted access to innovation, technologies, and resources. [2, 3]. A reliable system for forecasting future rice yield is required to achieve food security. However, because of the variability of the factors influencing rice output, developing a one-size-fits-all forecasting model is difficult. Moreover, the conventional practice often relies on historical data and expert recommendations and may not consider all factors influencing rice production in forecasting rice yields [17].

To overcome these challenges, machine learning approaches appear as recent alternatives to build prediction models for crop yields. It has a wider ability to capture complex relationships between various characteristics and may incorporate massive amounts of data, such as climate data. However, further study is needed in Malaysia on the utility of integrating weather-related information into machine learning models for rice crop prediction. In recent years, there has been a rise in interest in using machine learning approaches to create prediction models for crop output [2-4] and examine the influence of climate on agricultural productivity. Due to the continuous interaction of various variables impacting rice production, predicting rice yields using machine learning models can be challenging.

The climate is one of the impacting factors in agriculture, including rice. A previous study has revealed that climatic conditions influence rice production in Malaysia [2, 3]. However, it is unclear how much climatic data can increase the accuracy of rice crop estimates in Malaysia. This study investigates the effectiveness of including climate data in predicting rice production in Malaysia using linear regression. Our hypothesis was that including climate data in the prediction model could increase the accuracy of rice yield estimation in Malaysia, as climate is crucial to rice production. This study aims to provide insights into the possible benefits of integrating climate data for rice production prediction in Malaysia by including it in a regression model. In relation to this, the yearly rice yield information, as well as season indicators for the main and secondary plantation seasons of 10 years from states in Malaysia, were employed in the

modeling. In addition, climatic data as predictors such as wind speed, temperature, humidity, and rainfall were also included in the model.

This study is organized as follows: a complete overview of related studies, methodology, findings, and a discussion of the research's significance for agricultural practices and future research in Malaysia. The study is expected to contribute to the growing body of literature on the impact of climate change on agriculture in Malaysia, as well as give important insights for policymakers and farmers in this country.
.

2 Related works

Agriculture plays a crucial role in many countries, providing food and employment opportunities for millions of people. Accurate crop yield prediction is essential for farmers, policymakers, and other stakeholders because it can drive agricultural production, distribution, and pricing decisions. Traditional approaches to yield prediction have relied on statistical models and expert knowledge. However, recent advances in machine learning and data science have led to the development of more accurate predictive models for crop yields [5].

Rice yield prediction models can be modeled based on three approaches that are mechanistic, statistical/machine learning, and deep learning-based [19]. Regression modeling is a popular technique under statistical machine learning models for developing predictive models in agriculture. Regression models aim to establish a relationship between a dependent variable (in this case, crop yield) and one or more independent variables (such as climate data, soil quality, and agricultural practices). Regression models can be simple or complex, depending on the number and type of independent variables used. The quality of the data and the correlation among the variables used to generate the model might have an impact on regression performance. [6].

There has been growing interest in using regression models to predict crop yields with climate data recently. Climate factors such as temperature, humidity, rainfall, and wind speed are essential predictors of crop yields in many regions [7]. By incorporating climate data into regression models, researchers have developed more accurate and reliable predictions of crop yields [2, 3]. [18] replicated factors influencing rice production by combining typical independent variables such as temperature, precipitation, sunlight hours, and relative humidity to develop a deep learning-based rice yield forecast model. Other rice yield prediction models based on deep learning can be seen in. [19]

In Malaysia, machine learning in combination with climatic data has proved very useful for predicting rice yields. As a staple crop, thus forecasting rice harvests is critical for guaranteeing food security and economic stability [1]. In Malaysia, researchers discovered that adding climate data into regression models may greatly increase the

accuracy of rice yield predictions [2, 3]. Because climate involves many types of predictors, an experiment using a different climate predictor may yield a different result. This study emphasizes the need of using climate variables when developing forecast models for rice yields in Malaysia.

To summaries, regression modeling and the integration of climate information are important input for predicting crop yields in agriculture. Regression models may be used to create correlations between dependent and independent variables, and using climate data as an independent variable can enhance forecast accuracy. In the context of rice production prediction in Malaysia, adding climate data has shown to be a significant method for boosting forecast accuracy and guaranteeing food security in the country.

3 Methodology

This study's methodology section focuses on predicting rice yields in Malaysia using regression modelling techniques, with a particular emphasis on the role of climate data in this process. We employed a series of steps to achieve this, including data collection and analysis, data preparation, and regression modelling. Fig. 1 provides a flowchart of these steps and their interconnectedness in the overall process.

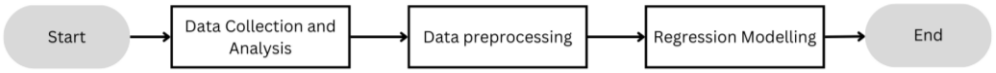

Fig. 1. Flowchart of the Steps Involved in Predicting Rice Yields Using Regression Modelling and Climate Data in Malaysia

3.1 Data Collection and Analysis

There are two types of datasets used in this study; firstly, the climate dataset and secondly is, rice production data. Both datasets were contributed by different agencies and were merged for mining.

The Malaysian Meteorological Department took the climate data used in this study. It consists of four numerical variables: wind speed, temperature, humidity, and rainfall. These data were collected from 2010 to 2021 for each month and are available for all states in Malaysia. These variables are essential in the study of agriculture as they affect plant growth and development, particularly in the case of rice. The suitability of these data for the study of agriculture has been shown in previous studies [2, 3].

The Department of Statistics Malaysia provided the rice production data. This dataset was explained by three numerical variables that are rice yield, parcel area, and planted

area. In addition, the data includes categorical variables such as the state of Malaysia, year, and season indicators. The rice yield data are reported annually for each state in Malaysia, with a range of values from 1.48 to 6.56 tons per hectare between 2011 and 2021. The parcel area represents the land used for rice cultivation, while the planted area represents the area where the rice crop is planted. These variables are necessary for prediction because they represent information on the quantity of land utilised for rice cultivation, which directly impacts yield. Adding categorical variables, including state, year, and season indicators, allows for examining how these variables impact the rice production forecast.

3.2 **Data preparation**

The data preparation stage is an important phase in data analytic studies since it includes transforming the raw data into a suitable format for mining. Several types of data preparation approaches were employed in this work to ensure the quality and accuracy of the data used in constructing the prediction model using regression. The data preparation has six tasks that involve missing data imputation, data transformation, data combining/merging, data scaling, one hot encoding, and data splitting. The flowchart in Fig. 2 depicts the order of these processes and how they are linked in the entire data preparation process.

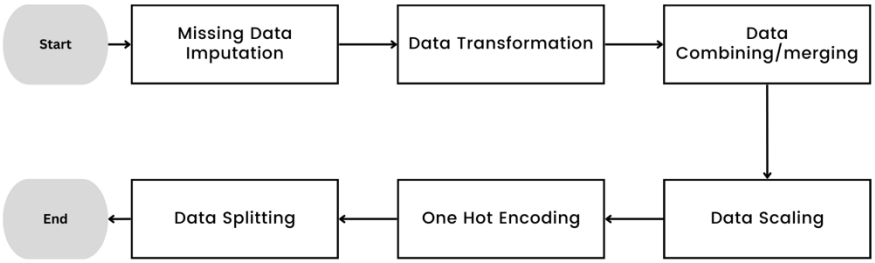

Fig. 2. Data preparation process for rice yields prediction system

The first task in data preprocessing is to solve the missing value problem. This problem is a common issue when using climate data for regression. Several reasons contribute to missing data, such as equipment failure or human error. Imputation is often used to fill in missing data to ensure the models are as accurate as possible. To estimate the missing value, a common imputation approach is to use the average value of the same month in the year before and after, as shown in equation (1). This solution can produce more accurate imputations than the overall average of accessible data [8–10]. While there are various ways of imputation, such as interpolation or machine learning algorithms, utilizing the average of the same month in the previous and subsequent years is a straightforward and effective strategy that may be used in several circumstances.

$$Climate_y = \frac{Climate_{y-1} + Climate_{y+1}}{2} \quad (1)$$

Another imputation method that can be used to replace mining values is by taking the average of the previous and next month's data or values from the same month in the previous and succeeding years also can be used as an imputation method. It is shown in equation (2). This approach has the advantage of identifying inter-monthly variability in weather patterns and can reduce the impact of seasonal trends. Although it provides precise and detailed missing value estimation, it relies on more data and calculations [11]. The study used the first approach as the imputation method.

$$Climate_m = \frac{Climate_{m-1} + Climate_{m+1}}{2} \quad (2)$$

The dataset was transferred into a yearly-based form to reduce the impact of outliers and obtain a more stable representation of the climate variables for each year. The transformed value is taken from the median of each year because the median is more robust to outliers. Aggregating monthly data into yearly data for climate variables has been widely implemented in agriculture-related research [12–14]. Each climatic variable was transformed separately. At the end of the process, four additional variables were constructed each year, indicating the median wind speed, temperature, humidity, and rainfall.

The climate data combines the production data using the states and year properties. Each state and year combination from the production data is merged with the corresponding climate data using the same properties. This process ensures that the climate data is aligned with the correct production data and can be used to accurately train and evaluate the regression models. The combined dataset is then used as the input for the regression models, aiming to predict the rice yield based on the climate properties. This approach allows for a more comprehensive analysis of the factors affecting rice yield. Besides that, the outcome can potentially provide insights into how climate properties can be managed to improve yield.

One of the requirements for a regression model's performance is that the dataset be produced in a specified scaled format. Numerical features were normalized using the Scikit-learn library's MinMaxScaler [15]. Normalization is a standard data transformation technique that scales numerical data to a fixed range (often between 0 and 1) to guarantee that each feature is given equal weight during model training. Meanwhile, the values of categorical features were converted to binary representation using a one-hot encoding method with the Pandas and Scikit-learn libraries. [15, 16]. Categorical variables are converted into binary vectors via one-hot encoding, with each category represented by its own binary feature. This method avoids the model assuming any ordinal link between the categories and ensures that each category is addressed independently during model training.

The final step in data preparation for modeling is to divide the data into training and testing folds. This work separated the dataset into train and test sections in a 70:30 ratio.

3.3 **Regression modeling**

The final step of the study is modelling. It involves the construction of a prediction model. This research employed multiple linear regression modelling to predict rice yield based on the provided data. For modeling, rice yield (state, year, season indicator, yield volume) and climate information wind speed, temperature, humidity, rainfall) were fed to a regression algorithm. Regression modelling aims to establish an association function between the dependent variable (yield volume) and the independent variables (climate data and rice yield information) and use these associations to produce accurate predictions. The model's performance was assessed using mean absolute error (MAE) and root mean squared error (RMSE). By comparing the performance of the two models, it is possible to learn that including climate data in the regression model enhances its predicted accuracy. Fig. 3 depicts the rice yields prediction model for this study.

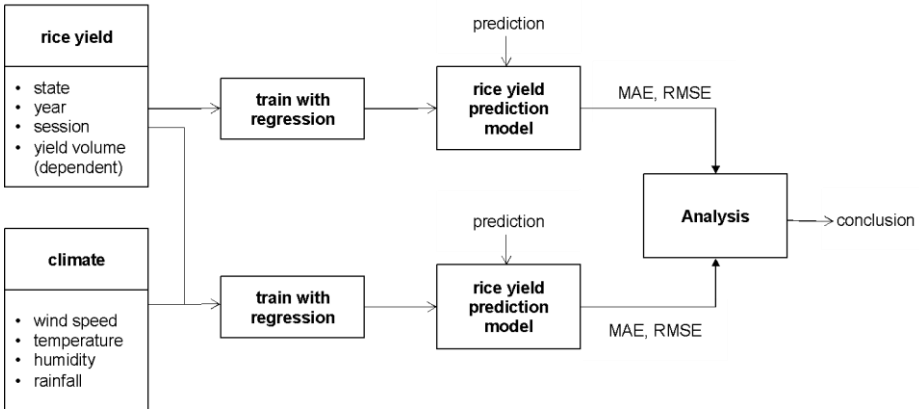

Fig. 3. Rice yields a prediction model based on multiple linear regression.

The improvement in prediction accuracy in a model can be quantified by comparing the model's error metrics before and after a modification or improvement is made. The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are commonly used error metrics for regression models. The percentage improvement in these metrics can be calculated using the following equation (3).

$$improvement(\%) = \frac{(PreviousError - UpdatedError)}{PreviousError} 100\%. \quad (3)$$

4 Results

This section outlines the study's findings. The aim is to reveal the impact of climate data on rice yield prediction in Malaysia using machine learning models. The experiment was divided into two stages. In the first stage, model performance is evaluated using only production data. In the second stage, models trained on both production and climatic data are evaluated. The mean absolute error (MAE) and root mean squared error (RMSE) were computed on the test set during the experiment. Both evolution metrics measure the prediction model's accuracy in predicting future rice yield. Lower error rates indicate a better model. The findings offer insights into the potential advantages of using climate data in rice yield prediction models and help instruct policymakers and farmers on better crop management techniques in the face of climate change. Table 1 depicts the rice yield prediction result using regression analysis.

Table 1. Performance Metrics for Rice Yield Prediction Model

Metrics	Production data only	Production and climate data	Improvements%
MAE	44612.60	10125.25	77.30
RMSE	58770.61	18059.56	69.27

The combination of climatic data in the prediction model enhanced the accuracy of the predicted rice yield significantly. The mean absolute error (MAE) decreased from 44,612.60 in the model that only used production data to 10,125.25 when climate data was incorporated. This represents a 77.30% decrease in MAE. Similarly, the root mean squared error (RMSE) decreased from 58,770.61 to 18,059.56, resulting in a 69.27% decrease in RMSE. These results demonstrate the importance of considering climate factors in predicting rice yield, as they can significantly improve the model's accuracy.

From the regression analysis, it can draw insights into how climatic data are used to forecast rice yields in the Malaysian states of Johor and Pahang. Fig. 4 demonstrates how including climate data considerably increased the regression model's accuracy, with the resulting regression line closely resembling the real data. However, the prediction without climatic data showed a significant offset from the actual data, with a difference of over 50,000. According to our study findings, climate factors like temperature and precipitation are critical in affecting rice yields in these states.

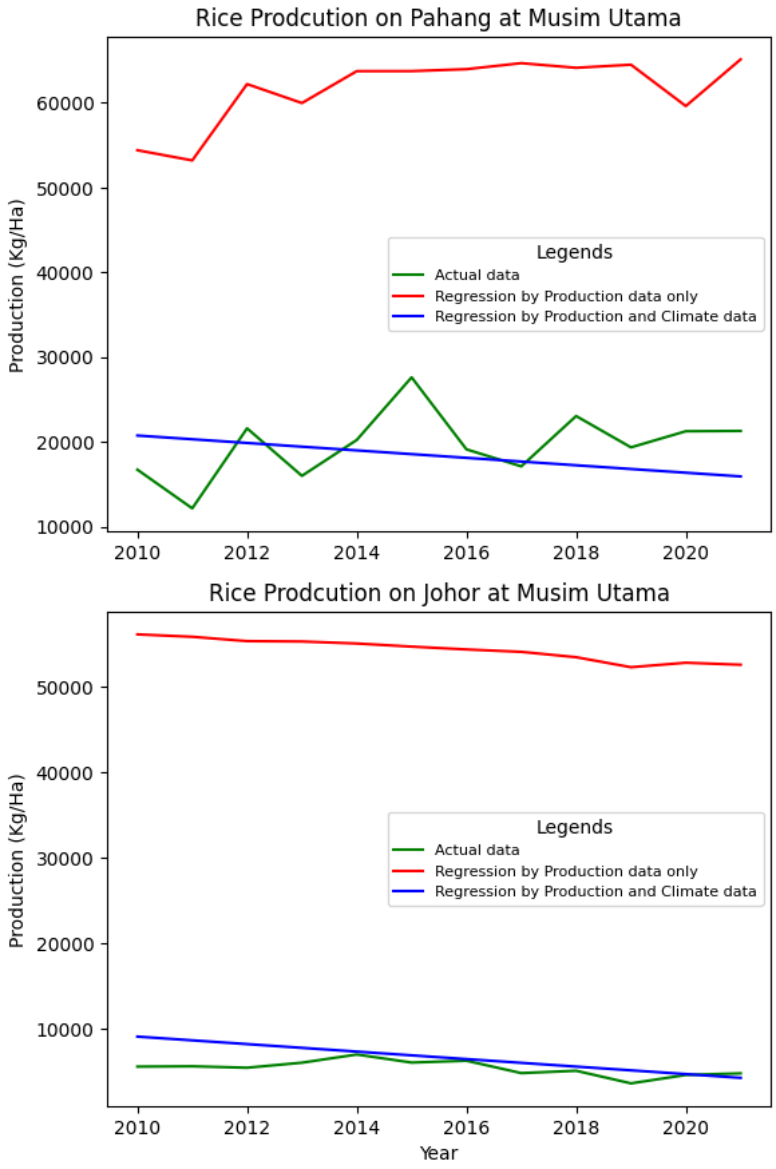

Fig. 4. Comparison of Rice Yield Predictions with and without Climate Data in Johor/Pahang, Malaysia.

Furthermore, our results demonstrate that the trend of rice production in Johor and Pahang is decreasing, albeit with varying degrees of decline. Specifically, the regression model for Johor predicts a decrease of approximately 14,000 metric tons of rice per year. In comparison, the regression model for Pahang predicts a reduction of

roughly 29,000 metric tons of rice per year. This suggests that rice production in both states faces significant challenges, and urgent measures are needed to address this issue.

These findings have important implications for policymakers and stakeholders, providing crucial information for designing effective strategies to improve rice production in these states. For instance, policymakers can focus on addressing the underlying factors contributing to the decline in rice production, such as changes in climate patterns and soil degradation. Furthermore, stakeholders can use these findings to develop more effective agricultural practices, such as using climate-resilient rice varieties and efficient irrigation systems.

In summary, our results demonstrate the importance of climate data in predicting rice yields in Malaysia and provide valuable insights into the trend of rice production in Johor and Pahang. These findings can inform evidence-based policymaking and stakeholder with the goal of increasing rice production and ensuring food security in the country.

5 Discussion

5.1 Conclusion

This work examines the impact of climate information on rice production forecasting in Malaysia. The reduction in MAE and RMSE by 77% and 69% demonstrated that incorporating climatic data greatly enhanced the model's accuracy. This shows that using climate data to improve the precision of models anticipating rice output could have significant ramifications for anyone involved in the agricultural sector, including farmers, decision-makers, and food distributors.

5.2 Limitations

It is important to acknowledge the limitations of this study. Firstly, the study's prediction model was limited to linear regression. Future research can investigate other machine learning techniques, such as random forests or neural networks, to further boost the model's accuracy. Second, Malaysia was the only nation using data in the study. To evaluate the generalizability of the findings, future studies can broaden the scope of the research to include other countries. Finally, because the study only included data from 2010–2021, it is possible that it did not fully account for the spectrum of climate variability that can affect rice production. Long-term studies that span a broader period may offer a more thorough understanding.

5.3 Future Studies

One of the potential improvements in future work is to employ advanced machine learning algorithms and experiment with other relevant variables that may affect rice production. For example, the soil types, irrigation systems, and insect control strategies that could affect rice production can be explored. The analysis can potentially be

broadened by including data from other areas to examine if the association between climate change and rice production is consistent across different regions. Another set of data points that can improve model accuracy and provide insight into the factors influencing rice production can be considered, such as socioeconomic characteristics, market prices, and governmental policies.

Acknowledgments

The authors of this article gratefully acknowledge the support provided by the Ministry of Higher Education (MoHE) through the Fundamental Research Grant Scheme (Ref: FRGS/1/2021/SS02/UUM/02/1 (S/O Code: 20107)). However, the views expressed in this article are those of the authors alone and do not necessarily reflect the official position of the MoHE, Malaysia.

References

1. Fatah FA (2017) Competitiveness and Efficiency of Rice Production in Malaysia. Georg-August-University Göttingen
2. Vaghefi N, Shamsudin MN, Radam A, Rahim KA (2013) Impact of Climate Change on Rice Yield in the Main Rice Growing Areas of Peninsular Malaysia. Res J Environ Sci 7:59–67. https://doi.org/10.3923/rjes.2013.59.67
3. Tan BT, Fam PS, Firdaus RBR, et al (2021) Impact of Climate Change on Rice Yield in Malaysia: A Panel Data Analysis. Agriculture 11:569. https://doi.org/10.3390/agriculture11060569
4. Sarr AB, Sultan B (2023) Predicting crop yields in Senegal using machine learning methods. International Journal of Climatology 43:1817–1838. https://doi.org/10.1002/joc.7947
5. Lobell DB (2013) The use of satellite data for crop yield gap analysis. Field Crops Res 143:56–64. https://doi.org/10.1016/j.fcr.2012.08.008
6. Montgomery DC, Peck EA, Vining GG (2021) Introduction to linear regression analysis. John Wiley & Sons
7. Ray DK, Ramankutty N, Mueller ND, et al (2012) Recent patterns of crop yield growth and stagnation. Nat Commun 3:1293. https://doi.org/10.1038/ncomms2296
8. Oriani F, Stisen S, Demirel MC, Mariethoz G (2020) Missing Data Imputation for Multisite Rainfall Networks: A Comparison between Geostatistical Interpolation and Pattern-Based Estimation on Different Terrain Types. J Hydrometeorol 21:2325–2341. https://doi.org/10.1175/JHM-D-19-0220.1
9. Junninen H, Niska H, Tuppurainen K, et al (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38:2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026
10. Nguyen V-H, Tuyet-Hanh TT, Mulhall J, et al (2022) Deep learning models for forecasting dengue fever based on climate data in Vietnam. PLoS Negl Trop Dis 16:e0010509. https://doi.org/10.1371/journal.pntd.0010509

11. Fassò A, Rodeschini J, Moro AF, et al (2023) Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy. Sci Data 10:143. https://doi.org/10.1038/s41597-023-02034-0
12. Rossi D, Mascolo A, Mancini S, et al (2023) Modelling and Forecast of Air Pollution Concentrations during COVID Pandemic Emergency with ARIMA Techniques: the Case Study of Two Italian Cities. WSEAS TRANSACTIONS ON ENVIRONMENT AND DEVELOPMENT 19:151–162. https://doi.org/10.37394/232015.2023.19.13
13. Toma MB, Belete MD, Ulsido MD (2023) Trends in climatic and hydrological parameters in the Ajora-Woybo watershed, Omo-Gibe River basin, Ethiopia. SN Appl Sci 5:45. https://doi.org/10.1007/s42452-022-05270-y
14. Boomgard-Zagrodnik JP, Brown DJ (2022) Machine learning imputation of missing Mesonet temperature observations. Comput Electron Agric 192:106580. https://doi.org/10.1016/j.compag.2021.106580
15. Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. the Journal of machine Learning research 12:2825–2830
16. McKinney W (2011) pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing 14.9:
17. Sujarwo, A. N. Putra, R. A. Setyawan, H. M. Teixeira, and U. Khumairoh, “Forecasting Rice Status for a Food Crisis Early Warning System Based on Satellite Imagery and Cellular Automata in Malang, Indonesia,” Sustainability, vol. 14, no. 15, p. 8972, Jul. 2022, doi: 10.3390/su14158972.
18. Ni, T., Han, X., Liu, F., He, X., & Ling, F. (2022). Research on Rice Yield Prediction Model Based on Deep Learning. Computational Intelligence and Neuroscience, 1922561. doi: 10.1155/2022/1922561
19. T. Chu and J. Yu (2020), “An end-to-end model for rice yield prediction using deep learning fusion,” Computers and Electronics in Agriculture, vol. 174, p. 105471. DOI: 10.1016/j.compag.2020.105471.