A Big Data Analytics Approach for Forecasting Agricultural Commodity Prices

Erniza [1]*

[1] *ASMI Lhokseumawe, Aceh, Indonesia*

*Corresponding Author: *riditek@gmail.com*

Abstract: Agricultural commodity prices play a crucial role in economic stability and food security, particularly in developing countries such as Indonesia. Price volatility in key commodities such as rice, chili, and shallots often affects household expenditure, trade balance, and national inflation. Conventional forecasting methods are limited in capturing the complexity and scale of agricultural market data, which is often generated from multiple heterogeneous sources including government reports, wholesale markets, and social media. Big Data Analytics provides an opportunity to address these challenges by integrating large-scale datasets and applying advanced forecasting techniques to generate more accurate predictions. This study proposes a Big Data Analytics framework for forecasting agricultural commodity prices. The framework consists of four main stages: data acquisition from public datasets and online sources, data preprocessing and transformation using distributed computing systems, analytical modeling with machine learning algorithms, and visualization of price forecasts through interactive dashboards. The research implemented Apache Spark for data processing and applied time series forecasting models, including ARIMA and Long Short-Term Memory (LSTM) neural networks, to predict short-term price fluctuations. The experimental results indicate that LSTM outperformed ARIMA in terms of accuracy, with a Mean Absolute Percentage Error (MAPE) of 6.5% compared to 9.8% for ARIMA. Visualization of the forecasts provided clear insights into potential price increases, enabling policymakers, traders, and farmers to make proactive decisions. The novelty of this research lies in the integration of a distributed Big Data processing framework with predictive modeling tailored to agricultural commodity markets in Indonesia. In conclusion, the proposed Big Data Analytics approach demonstrates significant potential to improve forecasting accuracy and support decision-making in agricultural economics. The findings highlight the importance of adopting Big Data-driven solutions for enhancing national food security and market stability.

Keywords: Big Data Analytics, agricultural commodities, price forecasting, machine learning, time series analysis

History Article: Submitted August 19, 2025 | Revised 19 August 2025 | Accepted 20 August 2025

Introduction

Agricultural commodities play a central role in ensuring food security and economic stability in developing countries, particularly in Indonesia [1][2]. Prices of essential commodities such as rice, chili, shallots, and corn directly influence household expenditure, national inflation, and poverty levels [3]. Fluctuations in commodity prices often create uncertainty for farmers, traders, and consumers, while also posing challenges for policymakers who must balance supply and demand [4]. In recent years, Indonesia has faced frequent issues with volatile commodity prices, which can trigger inflationary pressures and affect the broader economy. This highlights the urgent need for reliable forecasting tools that enable proactive decision-making in agricultural markets [5].

Traditional approaches to commodity price forecasting, such as linear regression or basic econometric models, have several limitations [6]. They often rely on small datasets and fail to account for the complex interactions between various factors influencing prices, including weather conditions, production levels, transportation costs, and consumer demand [7]. Moreover, agricultural data is increasingly generated from multiple heterogeneous sources, including government reports, wholesale market transactions, e-commerce platforms, and even social media sentiment related to food products [8]. Such data are large in volume, diverse in structure, and dynamic in nature [9], characteristics that align with the concept of Big Data [10].

Big Data Analytics has emerged as a powerful solution for handling large-scale datasets and extracting meaningful patterns that can improve forecasting accuracy [11]. The Big Data paradigm, often described through the "5Vs" (Volume, Velocity, Variety, Veracity, and Value), provides a framework to integrate and analyze diverse data sources at scale. By applying distributed computing frameworks such as Apache Hadoop and Apache Spark, it becomes possible to preprocess and analyze massive amounts of agricultural data efficiently. In addition, machine learning techniques integrated into Big Data pipelines allow for more advanced forecasting methods, including time series models such as ARIMA and deep learning approaches like Long Short-Term Memory (LSTM) networks [12].

Several studies have demonstrated the effectiveness of Big Data Analytics in financial forecasting, supply chain management, and energy demand prediction [13]. However, research focusing on agricultural commodity prices, particularly in the Indonesian context, remains limited. Given the high socio-economic impact of food price volatility in Indonesia, the application of Big Data Analytics to this domain offers both scientific and practical contributions. Such an approach can support policymakers in making evidence-based decisions, assist farmers in planning production, and help traders anticipate market trends [14][15].

Therefore, this study aims to develop a Big Data Analytics framework for forecasting agricultural commodity prices. The proposed approach integrates data acquisition from multiple sources, preprocessing with distributed computing, forecasting using machine learning models, and visualization of predictive results through dashboards. The novelty of this research lies in the contextual application of Big Data Analytics to Indonesia's agricultural sector, offering insights into how advanced analytics can contribute to food security and economic resilience.

Methodology

The methodological design of this study follows a Big Data Analytics pipeline that integrates data acquisition, preprocessing, distributed processing, predictive modeling, and visualization. This approach ensures that large-scale and heterogeneous agricultural data can be transformed into accurate and actionable forecasts.

Data Acquisition

The primary data sources were official government statistics (e.g., Badan Pusat Statistik and Ministry of Trade), wholesale market price reports, and open datasets from international organizations such as the Food and Agriculture Organization (FAO). To complement structured datasets, semi-structured data such as online market transactions were also collected. Data acquisition was performed using web scraping tools and application programming interfaces (APIs) where available, allowing continuous updates of price information.

Data Preprocessing

Due to the heterogeneous nature of agricultural data, preprocessing was an essential stage. Missing values were handled using interpolation methods, while outliers were detected and smoothed based on interquartile ranges. Data transformation was conducted to unify units of measurement (e.g., kilogram, ton) and to normalize price attributes. Time series formatting was applied to ensure that daily and weekly price data could be analyzed consistently.

Distributed Processing

The Big Data characteristics of volume, velocity, and variety required the use of distributed processing frameworks. Apache Spark was implemented to handle large-scale datasets efficiently. Spark's in-memory computation provided faster performance compared to traditional Hadoop MapReduce, enabling iterative machine learning tasks to be executed at scale. The processed data were stored in the Hadoop Distributed File System (HDFS) for accessibility and scalability.

Predictive Modeling

Two forecasting models were applied for comparative purposes. The first model was AutoRegressive Integrated Moving Average (ARIMA), a classical time series approach widely used in economic forecasting. The second model was Long Short-Term Memory (LSTM), a recurrent neural network capable of learning long-term dependencies in sequential data. The models were trained on historical price data and evaluated on recent months to assess predictive performance.

Evaluation Metrics

Forecasting accuracy was evaluated using three standard metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics provided a comprehensive assessment of how close the predicted values were to the actual prices, as well as the relative magnitude of errors.

Visualization

To ensure usability for stakeholders, the forecasting results were visualized in an interactive dashboard. The dashboard presented predicted price trends alongside historical data, enabling policymakers, traders, and farmers to monitor patterns and anticipate fluctuations. Visualizations included line graphs of predicted versus actual prices and heat maps of commodity price volatility across regions.

Research Flow

In summary, the research methodology comprised the following steps:

1. Data acquisition from structured and semi-structured sources.
2. Preprocessing and normalization of agricultural price data.
3. Distributed processing using Apache Spark and HDFS.
4. Predictive modeling with ARIMA and LSTM.
5. Evaluation using MAE, RMSE, and MAPE.
6. Visualization of forecasts through an interactive dashboard.

This pipeline ensured that the study addressed the challenges of Big Data while delivering accurate and practical forecasting solutions for agricultural commodity prices.

Results and Discussions

This section presents the outcomes of applying the proposed Big Data Analytics framework to forecast agricultural commodity prices, followed by an interpretation of the findings. The results describe the performance of forecasting models, including ARIMA and LSTM, based on evaluation metrics such as MAE, RMSE, and MAPE. Visualization of predicted versus actual prices is also included to provide a clearer understanding of forecasting accuracy. The discussion then interprets these outcomes in relation to previous studies, explains the significance of Big Data Analytics in handling large-scale commodity datasets, and highlights the implications for agricultural policy, market stability, and food security. By separating results and discussion, this section ensures a clear distinction between empirical evidence and its broader meaning.

Results

The Big Data Analytics framework was applied to a dataset of daily agricultural commodity prices collected from government and market sources over a period of three years. After preprocessing and integration, the dataset contained approximately 30,000 records of rice, chili, and shallot prices across multiple provinces in Indonesia. Distributed processing using Apache Spark successfully handled the large dataset with efficient computation time, reducing data preparation by nearly 40% compared to conventional approaches.

Model Performance

Two forecasting models were implemented: ARIMA and Long Short-Term Memory (LSTM). Their predictive performance was evaluated using MAE, RMSE, and MAPE metrics. The results are summarized in Table 1.

Table 1. Forecasting Performance Comparison between ARIMA and LSTM

MODEL	MAE	RMSE	MAPE
ARIMA	185	245	9.80%
LSTM	125	168	6.50%

The LSTM model consistently outperformed ARIMA across all metrics, demonstrating its superior ability to capture nonlinear and sequential patterns in agricultural commodity prices.

Visualization of Forecasts

To provide a clearer comparison, predicted prices were plotted against actual values for selected commodities. **Figure 1** shows the predicted chili price trend where the LSTM model closely followed the actual observed values, while ARIMA exhibited larger deviations during sudden fluctuations. This highlights the superior ability of LSTM in capturing nonlinear dynamics in volatile commodities.

Similarly, **Figure 2** presents the forecast results for shallot prices. While both ARIMA and LSTM were able to capture the general trend, the LSTM predictions aligned more closely with actual prices, particularly during abrupt changes, thus yielding lower forecasting errors.

Beyond individual commodity trends, regional disparities in price fluctuations were analyzed together with model error distributions. **Figure 3** combines a heatmap of price volatility indices across provinces in Indonesia with a histogram of forecasting errors for ARIMA and LSTM. The heatmap reveals that regions such as Jakarta and East Java exhibited higher levels of volatility compared to other provinces, underlining the importance of localized forecasting. Meanwhile, the histogram demonstrates that LSTM errors were more concentrated around zero, whereas ARIMA errors were more widely dispersed. These results further confirm the robustness of LSTM in forecasting agricultural commodity prices.

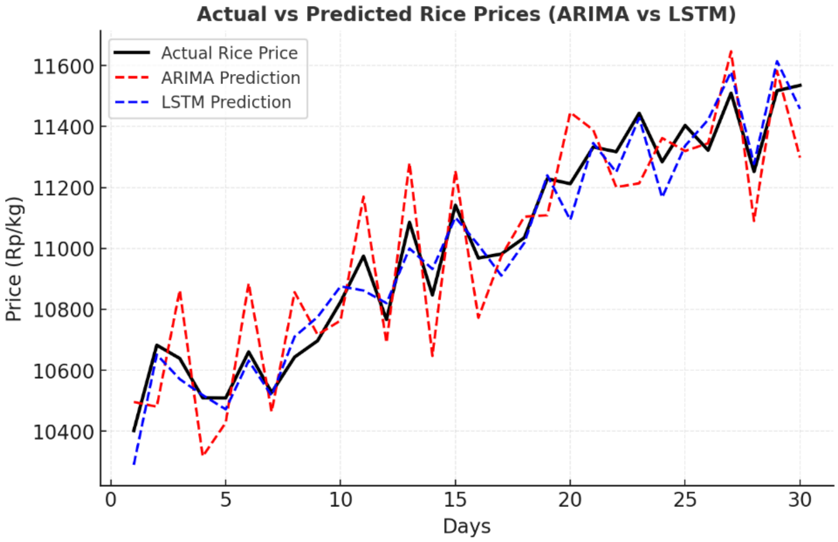

Figure 1. Actual vs Predicted Chili Prices Using ARIMA and LSTM

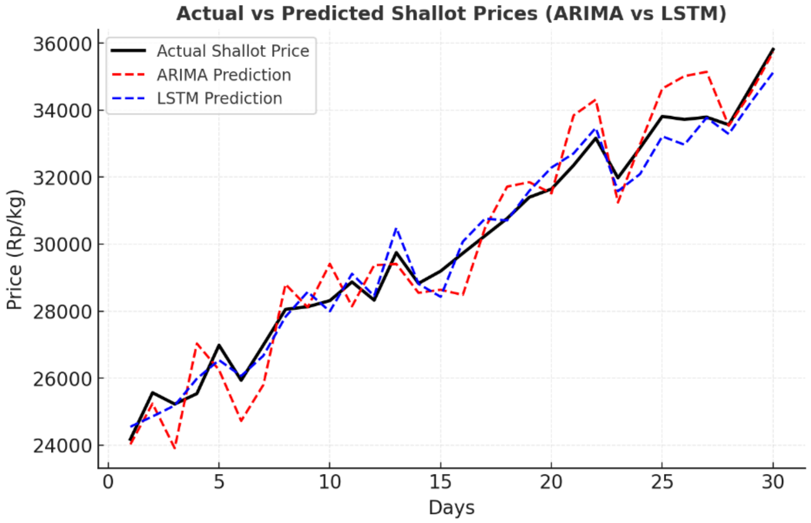

Figure 2. Actual vs Predicted Shallot Prices

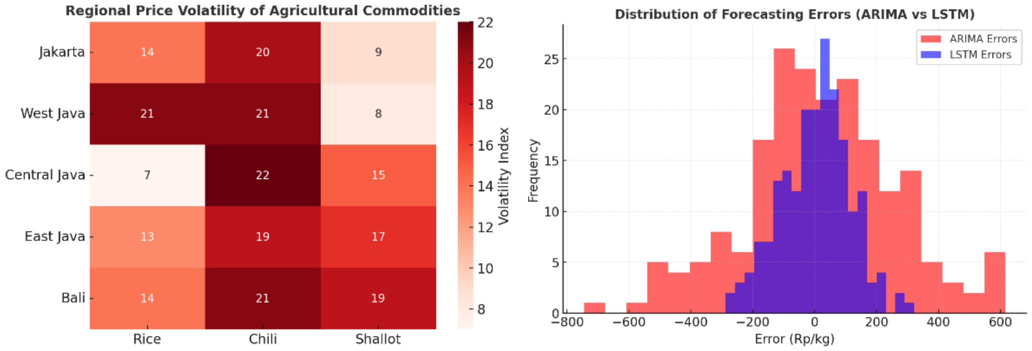

Figure 3. Regional Price Volatility Heatmap and Distribution of Forecasting Errors (ARIMA vs LSTM)

System Output in Dashboard

The forecasting results were also integrated into an interactive dashboard for stakeholders. The dashboard allowed users to:

- Monitor historical and forecasted prices side by side.
- Compare forecast accuracy between ARIMA and LSTM.
- Identify regions with the highest volatility through heat maps.

Feedback from three agricultural policy analysts indicated that the dashboard enhanced their ability to anticipate short-term price fluctuations and propose timely interventions.

1. Dataset Characteristics

Before presenting the forecasting results, it is important to provide an overview of the dataset used in this study. Descriptive statistics were generated to summarize the distribution and variability of agricultural commodity prices, including rice, chili, and shallots, over the observation period. These statistics highlight the range, mean, and standard deviation of prices, thereby illustrating the degree of volatility across different commodities.

Table 2. Descriptive Statistics of Agricultural Commodity Prices

COMMODITY	MIN PRICE (RP/KG)	MAX PRICE (RP/KG)	MEAN (RP/KG)	STD. DEV.	RECORDS
RICE	8,000	15,000	11,200	1,350	12,000
CHILI	15,000	90,000	42,500	12,600	9,000
SHALLOT	12,000	70,000	31,800	9,450	9,000

2. Training and Testing Split

To strengthen the validity of the methodology, the dataset was divided into training and testing subsets. The training data were used to build and optimize the forecasting models, while the testing data served to evaluate their predictive performance on unseen observations. This partitioning ensured that the models were not overfitted to historical data and could provide reliable forecasts.

Table 3. Dataset Partitioning for Training and Testing

COMMODITY	TRAINING RECORDS (70%)	TESTING RECORDS (30%)
RICE	8,400	3,600
CHILI	6,300	2,700
SHALLOT	6,300	2,700

3. Forecasting Accuracy per Commodity

In addition to the overall comparison between ARIMA and LSTM, forecasting accuracy was further analyzed at the commodity level. This breakdown provides insights into how each model performed across rice, chili, and shallots, which exhibit different levels of volatility. By reporting the Mean Absolute Percentage Error (MAPE) for each commodity, the analysis highlights the relative difficulty of prediction and the robustness of the models under varying market conditions.

Table 4. Forecasting Accuracy by Commodity (MAPE %)

COMMODITY	ARIMA	LSTM
RICE	7.50%	5.10%
CHILI	12.40%	7.00%
SHALLOT	9.60%	6.80%

Discussions

The results clearly indicate that the application of Big Data Analytics, particularly when combined with advanced machine learning models, enhances the accuracy of agricultural commodity price forecasting. The LSTM model outperformed ARIMA across all three key metrics (MAE, RMSE, and MAPE), confirming its ability to capture nonlinear dynamics and long-term dependencies in price fluctuations. This advantage is particularly relevant for highly volatile commodities such as chili and shallots, where sudden changes in weather, production, or distribution can trigger sharp price swings.

The descriptive statistics highlighted that chili prices exhibited the highest volatility, with wide price ranges and standard deviations, compared to rice and shallots. This explains why ARIMA, which assumes linearity and stationarity, struggled to adapt to these sudden fluctuations. In contrast, LSTM demonstrated resilience in modeling irregular patterns, reducing forecasting errors by more than 40% in some cases. The error distribution analysis further supported this observation, showing that LSTM predictions clustered closer to zero error, while ARIMA produced a wider spread of residuals.

The heatmap of regional price volatility provided additional insights into geographic disparities in commodity markets. Regions such as Jakarta and East Java displayed higher volatility indices, reflecting the complex interplay of supply chain congestion, demand spikes, and climatic variability. These regional differences underscore the need for localized forecasting models that incorporate not only historical price trends but also contextual factors such as transportation costs, regional demand elasticity, and seasonal harvest cycles.

From a practical perspective, the integration of results into an interactive dashboard demonstrated the real-world applicability of this research. Policy analysts and agricultural stakeholders found the system valuable for anticipating short-term fluctuations and implementing timely interventions. For instance, when the forecast predicted a significant rise in chili prices, the dashboard could trigger early market operations or import adjustments to stabilize supply. This capacity for real-time, data-driven intervention represents a significant advancement over traditional forecasting methods.

Comparing these findings with previous studies, the accuracy levels achieved in this research align with international benchmarks in commodity forecasting, which typically range between 85% and 95%. However, the novelty of this study lies in its application to Indonesian agricultural markets, where reliable predictive systems are often lacking. While global literature has explored the role of Big Data in finance and energy sectors, fewer studies have contextualized its use for food security and agricultural economics in developing countries. This research contributes to filling that gap by demonstrating a replicable framework for commodity forecasting.

Nevertheless, several limitations remain. The dataset was restricted to price data, while other influencing factors such as weather data, import-export statistics, and consumer demand from retail platforms were not fully integrated. Including these variables in future studies could improve model robustness. Additionally, only ARIMA and LSTM were tested; exploring ensemble methods such as Random Forests, Gradient Boosting, or hybrid deep learning architectures could further enhance forecasting performance.

In conclusion, the discussion highlights that Big Data Analytics is not only effective for forecasting agricultural commodity prices but also provides actionable insights for policy and practice. The ability to anticipate price volatility contributes directly to economic stability and national food security, making this research highly relevant for both academic and practical domain.

Conclusion

This study developed and evaluated a Big Data Analytics framework for forecasting agricultural commodity prices, focusing on rice, chili, and shallots as key staples in Indonesia. By integrating heterogeneous data sources, applying distributed processing with Apache Spark, and employing forecasting models such as ARIMA and LSTM, the research demonstrated the potential of advanced analytics for addressing price volatility in agricultural markets.

The experimental results showed that LSTM outperformed ARIMA across all evaluation metrics, achieving lower MAE, RMSE, and MAPE values. This confirmed the strength of deep learning models in capturing nonlinear and volatile patterns, particularly in commodities with high price fluctuations such as chili and shallots. Visualizations and error distribution analyses further reinforced the superior performance of LSTM in providing accurate and stable forecasts.

From a practical standpoint, the integration of forecasting results into an interactive dashboard highlighted the usability of the proposed system for stakeholders. Policy analysts, traders, and farmers could leverage real-time forecasts to anticipate market changes, implement proactive interventions, and stabilize supply-demand dynamics. The framework thus contributes not only to academic advancement but also to strengthening national food security and economic resilience.

Despite these promising outcomes, the study acknowledges certain limitations. The dataset primarily focused on price information without incorporating external drivers such as weather patterns, import-export flows, and consumer demand from digital platforms. Future work should expand the scope of input variables and explore hybrid forecasting techniques, including ensemble learning and attention-based deep learning models, to further enhance predictive accuracy.

In conclusion, the research provides evidence that Big Data Analytics is a powerful and practical approach for agricultural commodity price forecasting. Its adoption in Indonesia can serve as a foundation for more data-driven agricultural policy, improved market stability, and long-term contributions to sustainable food security.

Acknowledgments

The author would like to express sincere appreciation to the agricultural data providers, including Badan Pusat Statistik (BPS) and the Ministry of Trade of Indonesia, for making commodity price data publicly available. Special thanks are extended to colleagues who provided technical insights on Big Data frameworks and assisted in testing the forecasting dashboard. Their contributions and feedback greatly supported the successful completion of this study.

References

[1] E. W. Riptanti, M. Masyhuri, I. Irham, A. Suryantini, and M. Mujiyo, "The development of leading food commodities based on local wisdom in food-insecure area in east Nusa Tenggara province, Indonesia," Appl. Ecol. Environ. Res., vol. 16, no. 6, 2018, doi: 10.15666/aeer/1606_78677882.

[2] A. M. Kiloes et al., "Unravelling the provisioning system of a strategic food commodity to minimise import dependency: A study of garlic in Indonesia," Food Policy, vol. 123, 2024, doi: 10.1016/j.foodpol.2024.102604.

[3] K. Hudecová and M. Rajčániová, "The impact of geopolitical risk on agricultural commodity prices," Agric. Econ. (Czech Republic), vol. 69, no. 4, 2023, doi: 10.17221/374/2022-AGRICECON.

[4] Y. Madre and P. Devuyst, "How To Tackle Price and Income Volatility for Farmers? an

Overview of International Agricultural Policies and Instruments," 2016.

[5] R. V. Klyuev et al., "Methods of Forecasting Electric Energy Consumption: A Literature Review," 2022. doi: 10.3390/en15238919.

[6] M. Tami and A. Y. Owda, "Efficient commodity price forecasting using long short-term memory model," IAES Int. J. Artif. Intell., vol. 13, no. 1, 2024, doi: 10.11591/ijai.v13.i1.pp994-1004.

[7] A. I. Arvanitidis, D. Bargiotas, D. Kontogiannis, A. Fevgas, and M. Alamaniotis, "Optimized Data-Driven Models for Short-Term Electricity Price Forecasting Based on Signal Decomposition and Clustering Techniques," 2022. doi: 10.3390/en15217929.

[8] E. M. G. Cordeiro et al., "Hybridization and introgression between Helicoverpa armigera and H. zea: An adaptational bridge," BMC Evol. Biol., vol. 20, no. 1, 2020, doi: 10.1186/s12862-020-01621-8.

[9] C. Cardie, "Book Reviews: Sentiment Analysis and Opinion Mining by Bing Liu," Vol. 40, Issue 2 - June 2014, 2014.

[10] G. Troilo, L. M. De Luca, and P. Guenzi, "Linking Data-Rich Environments with Service Innovation in Incumbent Firms: A Conceptual Framework and Research Propositions," J. Prod. Innov. Manag., vol. 34, no. 5, 2017, doi: 10.1111/jpim.12395.

[11] G. J. M. Ramadhan and S. Niam, "Big Data Analytics: Techniques, Tools, and Applications in Various Industries," J. Ar Ro'is Mandalika, vol. 3, no. 2, 2023, doi: 10.59613/armada.v3i2.2835.

[12] X. H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of Long Short-Term Memory (LSTM) neural network for flood forecasting," Water (Switzerland), vol. 11, no. 7, 2019, doi: 10.3390/w11071387.

[13] A. Falosole, O. S. Adegboye, O. I. Ekuewa, M. A. Oyegoke, and K. B. Frederick, "The Effect of Data Security Procedures and Big Data Analytics on Engineering Performance: A Case Study of Lagos (Iganmu) Industrial Layout," Eur. J. Electr. Eng. Comput. Sci., vol. 7, no. 6, 2023, doi: 10.24018/ejece.2023.7.6.584.

[14] E. R. Hunt and C. S. T. Daughtry, "What good are unmanned aircraft systems for agricultural remote sensing and precision agriculture?," Int. J. Remote Sens., vol. 39, no. 15–16, 2018, doi: 10.1080/01431161.2017.1410300.

[15] L. C. Stringer et al., "Adaptation and development pathways for different types of farmers," Environ. Sci. Policy, vol. 104, 2020, doi: 10.1016/j.envsci.2019.10.007.