scientific reports

Check for updates

OPEN

Integrated analysis of meteorological conditions and agricultural yields in Indonesia using causal learning and intelligent clustering for climate change mitigation

Teny Handhayani[1]✉, Desi Arisandi[2] & Wasino Wasino[2]

Seasonal patterns strongly influence traditional agriculture in Indonesia; therefore, climate change is likely to have a significant impact on crop production. This study explores the dependency relationships between meteorological conditions and agricultural yields in Indonesia by integrating meteorological data with agricultural yield data. The datasets are collected from a wide range (2010 - 2024) at the district level. The meteorological data are obtained from 100 meteorological stations across Indonesia. The proposed approach employs the Peter–Clark (PC) algorithm to generate causal graphs and an Intelligent Kernel K-Means (IKKM) method to classify regions based on similarities in meteorological conditions and agricultural yields. IKKM is effective for mapping regions according to shared climatic and yield characteristics. This study examines five major agricultural commodities (cocoa, coffee, oil palm, cayenne, and paddy). The IKKM method successfully groups each dataset into three clusters, achieving an average Silhouette score of 0.35. The resulting causal graphs reveal dependency relationships between meteorological variables and crop yields. The dependent relationships indicate that rising temperatures are likely influence the declining yields of cocoa, oil palm, and paddy. Panel regression results indicate statistically significant (p-value ≤ 0.05) negative effects of temperature on agricultural land use and crop productivity. Minimum temperature significantly reduces cocoa land area, while minimum, maximum, and average temperatures negatively affect oil palm productivity. Average temperature also has a significant adverse impact on paddy productivity. These findings serve as a warning that climate change may directly and indirectly affect agricultural industries, with potentially severe impacts on key regions that contribute substantially to annual crop production.

Keywords Agricultural yields, Causal learning, Climate change, Clustering

Investigations into climate change have become a substantial subject. A study found that limited present adaptation and losses of the modern food bands dominate the global impact of climate change[1]. Climate change mitigation and adaptation are critical for the selection and management of crop varieties and livestock[2]. Several studies have established a positive linkage between meteorological conditions and agricultural drought, including evidence from Iran[3]. In the Oum Er Rbia River basin, agricultural drought closely follows meteorological drought variability[4]. Increasing temperatures exceeding the thresholds are harming crop yields across countries[5].

In machine learning, causal learning examines variable dependencies, commonly modeled as graphical structures in Bayesian Networks. Examples of Bayesian Network structure learning algorithms are the Peter-Clark (PC) algorithm[6,7], Greedy Equivalence Search (GES)[8], and Fast Causal Inference (FCI)[6]. The extension of these algorithms is developed to handle complex datasets, such as mixed data, e.g., the Copula PC algorithm[9] and Kernel Alignment PC algorithm[10]. The Bayesian Network structure learning model has been used to analyze causal-effect in some studies, for instance, climate data[11,12], air pollution[13,14] and health[15,16]. Clustering is

[1]Teknik Informatika, Universitas Tarumanagara, Jakarta 11440, Indonesia. [2]Sistem Informasi, Universitas Tarumanagara, Jakarta 11440, Indonesia. ✉email: tenyh@fti.untar.ac.id

a machine learning technique that groups similar instances, where intelligent approaches automatically identify the number of clusters and have been successfully applied in data analysis[17,18,19].

Indonesia lies between latitudes 6°N and 11°S, and longitudes 95°E and 141°E[20]. Indonesia experiences a tropical climate with dry (April–September) and wet (October–March) seasons, though regional variations arise from rainfall patterns, monsoon winds, topography, and distance from the sea. Western Indonesia generally receives higher rainfall, central regions are drier with extended dry seasons, and eastern Indonesia exhibits less predictable rainfall patterns. Indonesia shows marked regional climate variation, and seasonal changes strongly influence its traditional agriculture. Therefore, climate change possibly has a strong effect on crops. Studies have identified climate change in Indonesia, with rising temperatures and altered wind speed patterns in some areas[14,21,22]. Those studies are limited to investigating the changes in meteorological conditions. There is a noticeable lack of explanation of the effect of its change. Addressing this gap will help inform a better understanding of the effect of climate change in Indonesia.

This study encourages exploring the dependence relationships between meteorological conditions and agricultural yields in Indonesia. The research question addresses how to investigate the dependence relationships between climate change, through meteorological conditions, and agricultural yields in Indonesia. This paper proposes to use an integrated analysis of meteorological conditions and yields in Indonesia using a causal learning approach. This paper implements column-based data integration and the PC algorithm to generate causal graphs from datasets. It also applies the Intelligent Kernel K-Means (IKKM) to analyze the clusters of regions. Analyzing the dependence relationships of meteorological conditions and agricultural yields is beneficial for government and society to support sustainable agriculture and food security in the face of climate change. The research contribution is an integration model for analyzing dependence relationships among variables from different sources of datasets. This paper examines crop yields: coffee, cocoa, palm oil, paddy, and cayenne.

Methods

This paper uses two datasets: meteorological datasets and agricultural datasets. The meteorological datasets are obtained from public data of the Meteorology, Climatology, and Geophysics Agency in Indonesia. The raw data is available at https://dataonline.bmkg.go.id/dataonline-home. The meteorological data is collected from 100 meteorological stations across Indonesia. The dataset contains variables of date, minimum temperature (°C), maximum temperature (°C), average temperature (°C), humidity (%), rainfall (mm), sunshine duration (hours), maximum wind speed (m/s), and average wind speed (m/s). The agriculture datasets are collected from the Ministry of Agriculture of the Republic of Indonesia, where the raw data are available at https://bdsp2.pertanian.go.id/bdsp/id/lokasi. Those datasets are the final version and were obtained from the government official website, where the procedure for data collection is under the responsibility of the data provider. The agriculture data contains variables of year, city, the number of production (tons), land areas (hectares), and productivity (tons/hectares). The agriculture data are gathered at the district level in Indonesia. The information about datasets is displayed in Table 1.

This paper proposes to use machine learning methods for the integrated analysis of meteorological conditions and agricultural conditions. The stages contain data integration, causal graphs generation, and clustering. Figure 1 shows the methodology used in this paper. The datasets are time series data. The integrated data is a single table containing variables of meteorology and agriculture. The integrated data is then used as an input to generate causal graphs and region mapping. It implements the PC algorithm to generate causal graphs from datasets. The causal graphs are utilized to investigate the dependence relationships among variables. Intelligent Kernel K-Means is employed to cluster the regions. The clustering outputs the region mapping.

This paper is limited to analyzing the existing variables in datasets and assuming that there are no latent variables. The reason for using the PC algorithm is that this method is useful in generating a graph from a dataset represented by a partially directed acyclic graph (PDAG) that represents the dependent relationships among variables[7]. It is assumed that the dependent relationships between variables are the guide to understanding the influence of a variable on another. Previous studies have successfully applied the PC algorithm for learning causal graphs from time-series data[14,23,24]. The output of the PC algorithm facilitates visual analysis. The reason for applying the IKKM is that this method can cluster regions without determining the number of clusters, which reduces human intervention in the data analysis[25,26]. Human intervention might raise bias in data analysis. Moreover, Intelligent Kernel K-Means has the advanced ability to cluster nonlinearly separable data[27].

The PC algorithm is a Bayesian Network structure learning method that uses conditional independence tests to generate a causal graph[6]. A causal graph is a graph that represents the cause-and-effect relationships between variables in a dataset. A simple causal graph $X_1 \rightarrow X_2$ implies X_1 is a direct cause variable of X_2. X_i is a possible indirect cause variable of X_j only if there is a direct path from X_i to X_j.

Commodities	The number of cities	Duration (year)
Cocoa	333	2010 - 2022
Coffee	326	2015 - 2022
Oil Palm	172	2010 - 2023
Cayenne	445	2018 - 2024
Paddy	515	2010 - 2024

Table 1. The crop yields datasets.

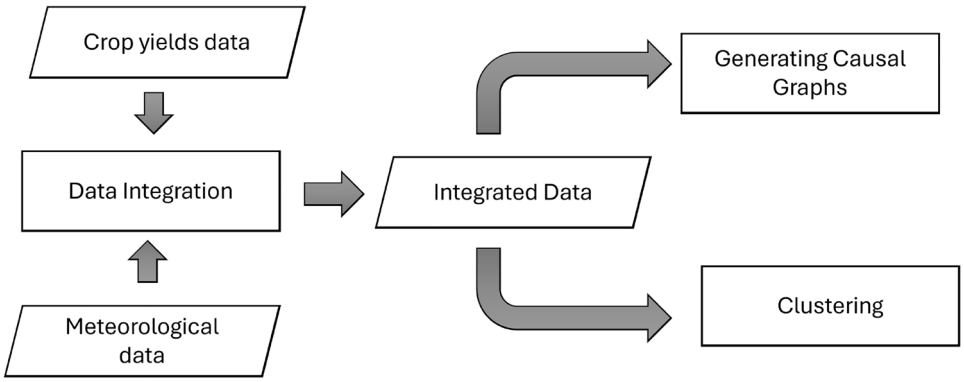


Fig. 1. The methodology in this paper.

The PC algorithm has two main steps: creating a graph skeleton and orienting the edges[7]. The conditional independence test is performed for every triplet of nodes. The first step of the PC algorithm produces a graph skeleton. These tests are then used to orient the edges. The PC algorithm is applied when it is assumed there are no latent variables in the dataset. A latent variable is a variable that is not recorded. The PC algorithm implements a conditional independence test to create a graph from a dataset. Let E, $\hat{\rho}$, α, n, and $\Phi(.)$ be the separation set, the partial correlation, the significance level, the number of samples, and the cumulative distribution function (cdf) of the normal distribution, respectively. The partial correlation can be calculated from the correlation matrix. A conditional independence test for Gaussian data can be explained using equation 1[9]. This equation tests 'is variable X_u conditionally independent X_v given X_E.

$$X_u \perp X_v | X_E \iff \sqrt{n - |E| - 3} \left| \frac{1}{2} \log \left(\frac{1 + \hat{\rho}_{uv|E}}{1 - \hat{\rho}_{uv|E}} \right) \right| \leq \Phi^{-1}(1 - \alpha/2) \tag{1}$$

Intelligent Kernel K-Means is inspired by Intelligent K-Means and begins by computing a kernel matrix using an appropriate kernel function (e.g., linear, RBF, or polynomial). This is a phase to transform the data from the input space into a high-dimensional feature space. The kernel matrix is then used as input for Intelligent Kernel K-Means[17]. Intelligent Kernel K-Means uses a procedure called an anomalous pattern to find centroids, then runs Kernel K-Means clustering. Let X, K, l, $\phi(S)$ be the dataset, the kernel matrix, the size of the kernel matrix, and the centre of mass, respectively. Suppose x and z are vectors in the kernel matrix, and i and j are index vectors in the kernel matrix. The procedures of the Intelligent Kernel K-Means can be explained as follows[26].

1. Compute a kernel matrix using a suitable kernel function, e.g., linear kernel: $K = XX^T$.
2. Compute a centre of mass (CoM) using equation $\phi(S) = \frac{1}{l} \sum_{i=1}^{l} \phi(x_i)$.
3. Compute the distance of all objects to the centre of mass using equation $\|\phi(x) - \phi(S)\|^2 = K(x,x) + \frac{1}{l^2} \sum_{i,j=1}^{l} K(x_i, x_j) - \frac{2}{l} \sum_{i=1}^{l} K(x, x_i)$.
4. Find the object C_1, that has the farthest distance to the centre of mass.
5. Compute the distance of all other objects to the C_1 using equation $\|\phi(x) - \phi(z)\|^2 = K(x,x) - 2K(x,z) + K(z,z)$.
6. Find the object C_2, that has the farthest distance to C_1.
7. Compute the distance of all objects to the initial centroids C_1 and C_2.
8. Group the objects into two clusters according to the nearest distance to the centroids. For instance, an object that has the closest distance to C_1 belongs to cluster S_1, and an object that has the closest distance to C_2 belongs to cluster S_2.
9. Run the iterated anomalous pattern to find other centroids.
   (a) For each cluster S_i, find the farthest object $C_i^{'}$ from the centroid in this cluster.
   (b) Choose a new centroid from the candidate centroids that has the maximum average distance to the existing centroids.
   (c) Run Kernel K-Means clustering[28].
   (d) The process of finding a new centroid will be terminated when there is no new centroid is found.
   (e) The clustering is terminated when there are no changes in cluster labels for each object.

Panel regression analysis was conducted to account for both cross-sectional and time-series variations in the data[29]. Fixed effects (FE) and random effects (RE) models were estimated to capture unobserved heterogeneity across entities. The fixed effects model (PanelOLS) controls for time-invariant characteristics by allowing individual-specific intercepts, while the random effects model assumes that unobserved effects are uncorrelated with the explanatory variables. To determine the most appropriate specification, the Hausman test was applied to compare the fixed and random effects models. The test results indicate whether the random effects estimator is consistent or whether the fixed effects model is more suitable for the analysis. Panel regression analysis was

employed to examine the relationship between meteorological conditions and crop yields using data that span multiple years and regions. The panel structure allows the analysis to capture both temporal variation over time and cross-regional heterogeneity. Crop yields were specified as the dependent variable, while meteorological variables—including temperature, rainfall, duration of sunshine, humidity, and wind speed—were included as independent variables.

The datasets are collected from data sources. The pre-processing data contains data cleaning, missing values handling, and data formatting. Data cleaning is dedicated to removing incorrect data. The type of missing data in the datasets is missing completely at random (MCAR), and the percentage of missing data is less than 20% for each variable. The missing values handling is conducted using forward and backward fill. The meteorological raw data is an annual time series of minimum temperatures, maximum temperatures, average temperatures, humidity, rainfall, sunshine duration, maximum wind, and average wind. Meanwhile, the agricultural data is annual for land area, production, and productivity. Hence, we compute the average of annual meteorological data. After the pre-processing step, the clean data is then used as input for column-based data integration.

This paper uses the PC algorithm from the R package *pcalg*[30]. This paper applies the PC algorithm that implements the conditional independence for Gaussian data (*gaussCItest*). The authors develop their own code in Python for Intelligent Kernel K-Means. The experiments run in Google Colab. The data formatting is changing the format of the data and making it suitable for the applications. Figures S1 (A), S2 (A), S3 (A), S4 (A), and S5 (A) are generated using open source Python Libraries Geopandas[31], and Cartopy[32]. The source maps were obtained from GADM, where they are freely available for academic and other non-commercial use[33].

Results and discussion

This paper investigates five crops: cocoa, coffee, palm oil, cayenne, and paddy. They are produced in specific regions; therefore, the number of producing regions differs by crop. These commodities play an important role in the economy of Indonesia, which can be explained as follows. Indonesia is one of the largest coffee bean producers, and the government encourages increasing the coffee bean production[3435]. Cocoa farming is found in many regions in Indonesia on different scales[36]. Rice is produced from paddy seeds, and it is the staple food for Indonesians. Chili and cayenne are widely used in Indonesian cuisine, with fresh cayenne commonly consumed as a snack accompaniment. In 2021, around 59% of the world's crude palm oil production was produced in Indonesia[37]. Indonesia has been one of the largest cocoa producers in the world[38].

The graphs are generated from datasets using the PC algorithms. First of all, it computes the distance of the cities of producers of the commodity to the meteorological stations and chooses the closest station. The next step is to implement column-based data integration to merge meteorological data from the nearest station and agricultural data. The integrated data is then used as input for the PC algorithm to generate a causal graph. Each commodity is processed individually. Note that the abbreviations can be found here: Minimum Temperatures (TN), Maximum Temperatures (TX), Average Temperatures (TA), Humidity (RH), Rainfall (RR), Sunshine (SS), Maximum Wind Speed (FX), Average Wind Speed (FA), Production (PD), Productivity (PV), and Land Area (LA). The causal graphs visualize the dependent relationships among variables in a dataset. The analysis highlights relationships between meteorological variables and agricultural variables. The causal graphs are displayed in Fig. 2.

The integrated data is also used for clustering analysis. The clustering analysis is intended to map areas according to the meteorological conditions and crop yields of the observed commodities. It implements the Intelligent Kernel K-Means using a Linear kernel function. The reason for implementing the IKKM algorithm is the ability to cluster a non-linearly separable dataset without determining the number of clusters that will be formed[27]. IKKM automatically finds the clusters, so it reduces the human intervention in the data analysis, which may lead to bias. First, the data is normalized to reduce the gap in values, and then it is clustered using IKKM. Intelligent Kernel K-Means automatically finds the number of clusters through an anomalous pattern procedure. The performance of Intelligent Kernel K-Means is compared to the K-Means to find out how well this algorithm works. The summary of clustering results and its evaluation is displayed in Table 2. The results show that IKKM produces higher average Silhouette scores and lower Davies-Bouldin Index than K-Means. It ensures that Intelligent Kernel K-Means has slightly better performance than K-Means. Note that all visualization of mapping Indonesian regions is created using Python. To support the analysis, it implements the panel regression (Random Effect (RE) test, Fixed Effect (FE) test, and Hausman test). The results of RE test, FE test, and Housman test for each dataset are displayed in Table S1, S2, S3, S4, and S5.

Cocoa

In the cocoa dataset, the maximum, mean, and minimum distances from cities to the nearest meteorological station are 363.67 km, 67.46 km, and 0.58 km, respectively. Only four cities have distances greater than 200 km, and none of them are major cocoa producers.

The investigation of dependence relationships between meteorological variables and cocoa crop yields is displayed in Fig. 2 (A). A causal graph from the cocoa dataset is generated using $\alpha = 0.05$. Maximum temperatures have a dependent relationship to cocoa production. The average and minimum temperatures are indirectly connected to cocoa production, which indicates that they may influence cocoa production. Rainfall, average wind speeds, and maximum wind speeds are likely to affect cocoa crop productivity. Average wind speed is a variable that has a dependent relationship to the land area of cocoa farming. Wind speed, rainfall, and sunshine duration show dependent relationships with cocoa productivity. Related studies find that meteorological variables, i.e., temperatures, humidity, rainfall, sunshine, and wind speed, influence cocoa yields[39]. Wind participates as one of the factors on the natural pollination of cocoa[40].

Figure S1 (A) shows the mapping areas of cluster analysis. Intelligent Kernel K-Means groups the cocoa farming areas into 3 clusters according to the meteorological conditions and crop yields. Cluster 2 contains

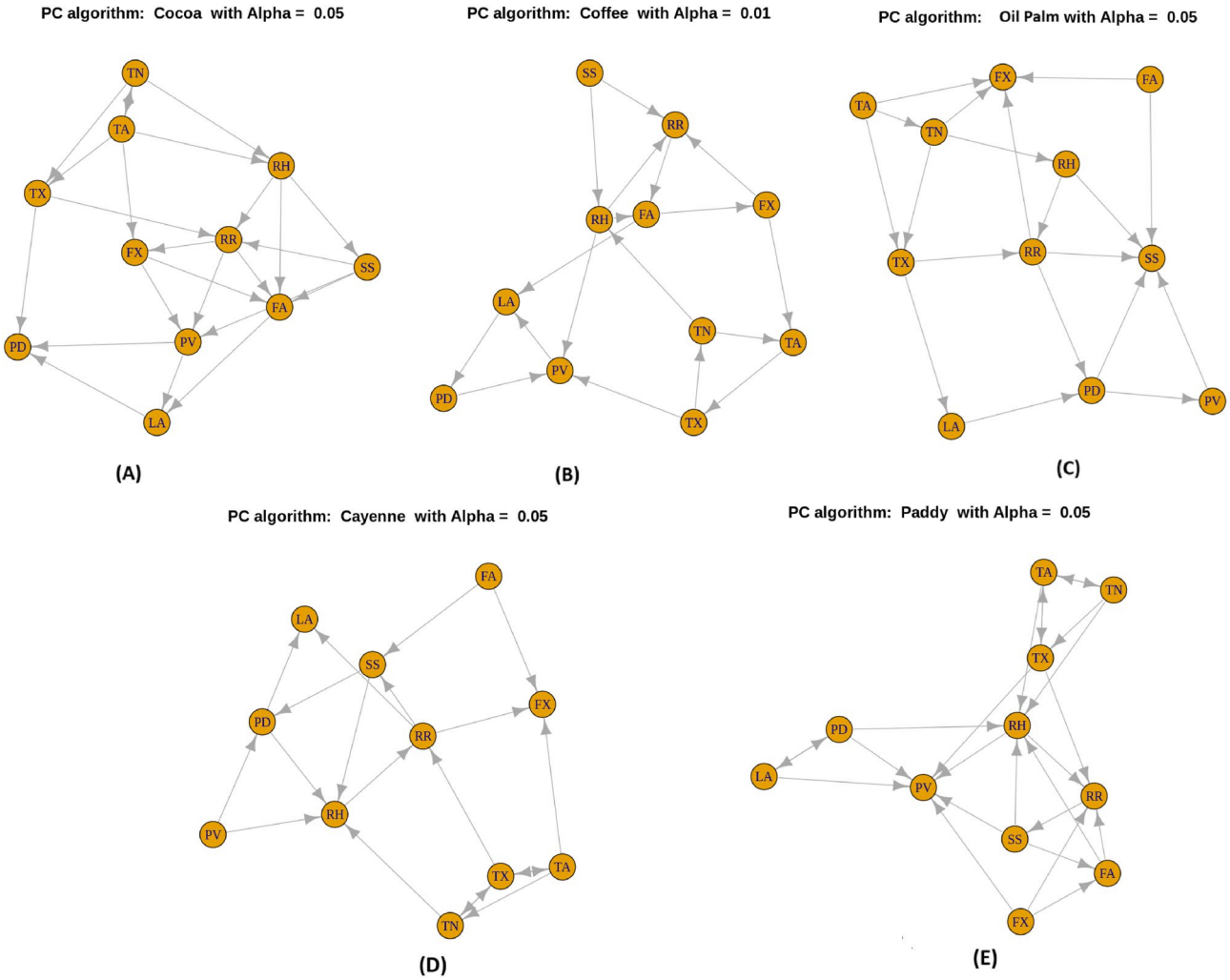


Fig. 2. Causal graphs between meteorological variables and crop yields: (**A**) Cocoa, (**B**) Coffee, (**C**) Palm Oil, (**D**) Cayenne, and (**E**) Paddy.

Commodity	Num. of Clusters	IKKM		K-Means	
		Avg. Silhouette	DBI	Avg. Silhouette	DBI
Cocoa	3	0.3737	1.0333	0.1480	1.9287
Coffee	3	0.4022	0.9585	0.2599	1.5517
Oil Palm	3	0.2874	1.1386	0.1534	2.0972
Cayenne	3	0.3913	1.1322	0.2369	1.6656
Paddy	3	0.2922	1.3704	0.1772	1.8400

Table 2. Clustering Evaluation Using Intelligent Kernel K-Means and K-Means Algorithm.

six districts, i.e., Parigi Moutong (Central Sulawesi), Luwu (South Sulawesi), Luwu Utara (South Sulawesi), Kolaka Utara (Southeast Sulawesi), Polewali Mandar (West Sulawesi), and Mamuju (West Sulawesi). These regions consistently achieve higher annual yields with Kolaka Utara, Parigi Moutong, Polewali Mandar, and Luwu Utara jointly contributing approximately 21.21% of the national cocoa output. The 95th percentile of maximum temperatures, approximately 34°C, was used to define extreme heat events. In the cities that are major cocoa producers, the years with the highest frequency of days exceeding this threshold were 2023, 2012, and 2016–2017, with 39, 25, and 27 days, respectively. There exist 36 cities in Cluster 1, and the rest are in Cluster 0. Cluster 1 is the areas that have a slightly higher cocoa crop than areas in Cluster 0. The areas in Cluster 1 are spread across parts of Java Island, Sulawesi Island, and others. The annual trend of minimum, maximum, and average temperatures has risen. The annual trend of minimum temperatures has slightly increased (see Figure S1 (B)). The minimum temperatures over the years have consistently increased. The detail RE test, FE test, and

Hausman test results for the cocoa dataset can be seen in Table S1. Random effects regression estimates that the minimum temperature has a negative and significant impact on cocoa land area. The random effect test also shows that the duration of sunshine has a negative coefficient on land area. A negative coefficient indicates that increasing minimum temperature and increasing sunshine duration tend to decrease land area, although the model's explanatory power is relatively low. It is a sign that the continuous increase in minimum temperatures harms cocoa farming. There is a decreased annual trend of maximum wind speed. An analysis of cocoa crop yields from 2010 to 2022 shows a decreasing trend in production and land areas. From 2014 to 2016, the cocoa productivity was increased. However, from 2016 - 2022, it has significantly decreased. The random effect test result shows that maximum wind speed has a positive and significant effect on production. It means that the increasing maximum wind speed is probably to increase cocoa production. If the trend of maximum wind speed has continuously decreased, it is likely to harm cocoa production. These results reinforce the understanding that cocoa cultivation requires a balanced combination of temperature, rainfall, and humidity. The persistent increase in temperature may have a negative influence on cocoa growth, consistent with previous findings that excessive heat and prolonged sunshine negatively affect cocoa yield[39].

Coffee

In the coffee dataset, the minimum, maximum, and mean distances from cities to the nearest meteorological station are 0.58 km, 356.71 km, and 66.5 km, respectively. Only two cities have distances greater than 200 km, and neither is a major coffee producer.

The dependence relationships between meteorological conditions and coffee yields in Indonesia are explained below. Figure 2 (B) displays dependence relationships between meteorological conditions and coffee crop yields created by the PC algorithm at $\alpha = 0.01$. A direct arrow exists from the average wind speed to land areas, which means that there is a dependent relationship between those variables. Humidity and maximum temperature are likely influencing variables for coffee productivity. The previous study shows that climate change affects Arabica coffee by increasing temperatures, altering rainfall patterns, and causing extreme weather events, which lead to increased disease incidence, post-harvest losses, shifts in suitable growing areas, and the expansion of khat and commercial trees[41]. Figure S2 (A) shows the mapping areas for cluster analysis of the coffee dataset. Mapping areas of coffee bean producers in Indonesia using Intelligent Kernel K-Means obtains three clusters. Cluster 0 contains Muara Enim, Lahat, OKU Selatan, Empat Lawang, Lampung Barat, and Tanggamus. The districts in Cluster 0 have higher coffee land areas and production. Cities in Cluster 0 have range of temperatures around 23°C to 33°C with average temperature from 26°C - 28°C. The maximum wind speed in Cluster 0 is around 5 to 7 m/s, with an average wind speed is around 2 m/s to 3 m/s. Areas in Cluster 0 have lower sunshine duration than areas in other clusters, which is about 4.5 to 5.8 hours. The threshold for extreme hot days is defined by the 95th percentile, corresponding to temperatures above 33.4°C. In cities that are major coffee producers, the years with the highest number of extreme hot days were 2024, 2023, and 2019, with 43, 34, and 30 days, respectively. From 2015 to 2022, the trend of coffee bean production, land areas, and productivity in Cluster 0 has increased. Cluster 1 contains 38 areas with the coffee bean productivity lower than that in Cluster 0 but higher than that in Cluster 2. Areas in Cluster 1 have temperatures around 19°C to 29°C. The remaining areas are collected in Cluster 2 with the lower coffee land areas, production, and productivity. The areas in Cluster 2 have a range of temperatures similar to the areas in Cluster 0. In general, there is no specific change in the annual trend of meteorological conditions in the areas of coffee bean production. There is no significant change in the annual trend of minimum, maximum, and average temperatures in Cluster 0, Cluster 1, and Cluster 2. For further analysis, it implements the RE test, FE test, and Hausman test, which are displayed in Table S2. Based on the results of the Random Effects estimation, humidity has a positive and significant effect on coffee land areas. On the other hand, the estimation results using the Random Effects model show that sunshine duration has a negative and significant effect on coffee land areas. It means that the increasing of sunshine duration likely reduces land areas. The PanelOLS estimation results with entity fixed effects show that average wind speed has a negative and significant effect on coffee production. This finding indicates that a decrease in average wind speed is correlated with an increase in coffee production.

Oil palm

In the oil palm dataset, the minimum, maximum, and mean distances from cities to the nearest meteorological station are 0.58 km, 230.05 km, and 78.68 km, respectively. Only one city has a distance greater than 200 km, and it is not a major oil palm producer.

The dependence relationships between meteorological conditions and oil palm crop yields can be found in Figure 2 (C). It is output by the PC algorithm at $\alpha = 0.05$. Maximum temperatures possibly influence the oil palm land area. Rainfall likely affects oil palm production. Minimum temperature, average temperature, and humidity are possibly indirectly influencing the oil palm crop yields. Previous research found that high temperatures and heavy rainfall significantly influenced palm oil production on the western coast of Sabah, Malaysia[42]. Mapping clusters of the oil palm dataset can be found in Figure S3 (A). Intelligent Kernel K-Means divides the oil palm dataset into three clusters. Cluster 0 contains regions Labuhan Batu, Asahan, Indragiri Hilir, Pelalawan, Siak, Kampar, Rokan Hulu, Bengkalis, Rokan Hilir, Musi Banyu Asin, Ketapang, and Kota Waringin Timur. Those regions have the largest average of oil palm production, land areas, and productivity. In the oil palm dataset, the threshold for extreme hot days is defined by the 95th percentile, corresponding to maximum temperatures above 33.4°C. In cities that are major oil palm producers, the three years with the highest numbers of extreme hot days were 2024, 2016, and 2019, with 43, 39, and 28 days, respectively. Cluster 0 has a range of temperatures around 23°C-33°C and humidity about 76% - 85%. From 2010 to 2023, the annual trend of sunshine duration in Cluster 0 has increased, but the annual trend of maximum and average wind speed has decreased. From 2010 to 2017, the annual trend of oil palm crop yields increased, but after that, it decreased.

Cluster 1 comprises 156 cities primarily located in parts of Sumatra, Kalimantan, and Sulawesi. Cluster 1 is the second rank of oil palm crop yields. Cities in cluster 1 have a similar range and trend in temperatures to areas in Cluster 0. Cluster 0 and Cluster 1 also have an increasing annual trend of sunshine duration and decreasing wind speed. The annual trend of minimum temperatures has slightly increased in all Clusters. The results of the random effect, fixed effect, and Hausman tests can be found in Table S3. The estimation results of the random effect test and fixed effect test show that all temperature measurements (Minimum, Maximum, and Average Temperature) have a negative and significant effect on oil palm productivity. Furthermore, the estimation results of those tests indicate that other variables, i.e., humidity, rainfall, and sunshine duration, have a negative and significant effect on oil palm productivity. This means that increasing temperatures, humidity, rainfall, and sunshine duration tend to decrease oil palm productivity. The random effect and fixed effect test results suggest that the increasing sunshine duration might decrease the oil palm production and land areas.

Cayenne

In the Cayenne dataset, the minimum, maximum, and mean distances between cities and meteorological stations are 0.53 km, 278.39 km, and 63.9 km, respectively. There are five cities with distances exceeding 200 km. These cities are located in eastern Indonesia, and none of them is a major cayenne producer.

A causal graph representing the dependence relationships between meteorological conditions and cayenne crop yields is displayed in Figure 2 (D). This graph is generated at $\alpha = 0.05$. There is a dependent relationship between sunshine duration and cayenne production, as well as between rainfall and cayenne land area. These relationships indicate that sunshine duration likely influences cayenne production, while rainfall may influence the extent of cayenne land area. The previous study found that rainfall contributions in the cultivation of chili pepper[43]. Maximum temperature and average wind speed exhibit indirect dependency relationships with cayenne production and land area. This suggests that both maximum temperature and average wind speed may influence cayenne production and land area. Previous research also showed that increasing temperatures decreased the cayenne production[44]. Mapping areas according to the meteorological conditions of Cayenne crop yields is displayed in Figure S4 (A). Intelligent Kernel K-Means groups cities of cayenne producers into three clusters. Cluster 2 contains 9 cities (Garut, Tuban, Sampang, Lombok Timur, Parigi Moutong, Kediri, Blitar, Malang, and Probolinggo) that have the highest production, land areas, and productivity of cayenne in Indonesia. Areas in cluster 2 has a minimum temperature around 22°C to 23°C, maximum temperatures around 30°C-32°C and average temperatures of 25°C - 27°C. Cluster 2 is characterized by the lowest rainfall trend and the highest sunshine duration among the clusters. Cluster 1 contains 40 cities with the second-highest cayenne productivity. Areas in Cluster 1 are mostly located on Java Island, parts of Sulawesi Island, and the Nusa Tenggara Islands. The remaining areas in Cluster 0 have lower cayenne crop yields. Cluster 0 has an average highest temperature than cities in other clusters. Meanwhile, cities in cluster 1 have the lowest temperatures and the highest rainfall. The annual trends of those variables are represented in Figure S4 (B). The annual trends of minimum, maximum, and average temperatures slightly increase in the areas of the cayenne producer. Extremely hot days are defined as the maximum temperatures exceeding the 95th percentile. The three years with the highest number of hot days in areas with major cayenne crop yields are 2023 (47 days), 2024 (39 days), and 2018–2019 (20 days). From 2020 to 2023, the annual trend of cayenne land areas, production, and productivity in cluster 2 has decreased. In Cluster 2, from 2018 to 2024, the trends in cayenne production, land area, and productivity have all decreased. It needs a proper investigation to study this issue. For further analysis, the random effects model, fixed effects model, and Hausman test were conducted. The results are presented in the Table S4. Although there is insufficient statistical evidence to conclude that meteorological conditions affect cayenne yields, the declining cayenne crop yields alongside rising temperatures remain a concern.

Paddy

In the paddy dataset, the minimum, maximum, and mean distances of cities to the nearest meteorological station are 0.53 km, 230.05 km, and 62.43 km, respectively. There is one city have distance of over 200 km, but it is not a major paddy producer.

A causal graph depicting the relationships between meteorological conditions and paddy crop yields is shown in Figure 2 (E). It is generated at $\alpha = 0.05$. In general, all meteorological variables (temperatures, humidity, rainfall, duration of sunshine, and wind speed) possibly affect paddy productivity. The influence of meteorological variables is likely direct and indirect. Humidity, sunshine duration, and maximum temperature are possibly influencing paddy productivity. Rainfall, average temperatures, average wind speed, and minimum temperatures are likely to affect paddy productivity indirectly. Figure S5 (A) shows a mapping of areas for the paddy dataset. Meanwhile, Figure S5 (B) displays the annual trend of meteorological variables and paddy crop yields. Cluster 0 comprises 37 districts with higher paddy production, land areas, and productivity. Those districts are spread out mainly in Java Island, and small parts of Sumatra, Kalimantan, and Sulawesi Island. In the paddy dataset, the threshold for extreme hot days is defined by the 95th percentile, corresponding to maximum temperatures above 33.4°C. In cities with major paddy production, the three years with the highest numbers of extreme hot days were 2023–2024 (45 days), 2019 (37 days), and 2015 (27 days). In general, the trend of annual minimum, maximum, and average temperature in the areas of paddy producers is increasing. From 2016, the trend of paddy production and land areas has decreased. There is a negative correlation between temperatures and paddy crop yields. Previous studies found that high temperatures surpassing the critical threshold possibly damage rice grain yields[45,46,47]. Maximum wind speed also has a negative correlation with paddy crop yields. Sunshine duration has an increasing trend over the years. Proper sunlight can increase paddy production. Indonesia is a tropical country where sunrise and sunset remain the same over time all year. The short sunshine duration usually happens in the wet season when the sunlight is covered by clouds. It means that continuously rising sunshine duration means that the cloud that brings rainfall is decreased. Paddy farming in Indonesia is

influenced by rainfall. From 2010 to 2024, the annual trend of rainfall has no specific trend. From 2010 to 2024, temperatures and sunshine duration are meteorological variables that have continued to increase. On the other hand, the annual trend of wind speed has continued to decrease. Table S5 shows the random effect test, fixed effect test, and Hausman test for the paddy dataset. The results of the random effects estimation indicate that all temperature measures (minimum, maximum, and average) have a negative and significant effect on paddy production, as well as on paddy land area. This suggests that increases in temperature tend to reduce both paddy production and the extent of paddy land. The results from both the Random Effects and PanelOLS Fixed Effects models indicate that average temperature has a negative and significant impact on paddy productivity. This suggests that increases in average temperature tend to reduce rice productivity. The estimation results from both the random effects and PanelOLS models with entity fixed effects consistently indicate that wind speed measurement (maximum and average) have a positive and significant impact on paddy production and land areas. This suggests that increases in maximum and average wind speed tend to enhance paddy production and expand cultivated areas.

From 2010 to 2024, Indonesia's meteorological conditions can be summarized as follows: the minimum, maximum, and average temperatures are approximately 23.65°C, 31.67°C, and 27.10°C, respectively. Average humidity is around 82.33%, average rainfall is 8.87 mm, and average sunshine duration is 5.26 hours. Maximum and average wind speeds are 4.76 m/s and 2 m/s, respectively. Using the 95th percentile, the threshold for hot days is approximately 33.19°C. Figure S6 shows the annual trend of meteorological condition in Indonesia from 2010 to 2024. The minimum and average temperatures show a consistent upward trend. Meanwhile, maximum and average wind speeds show a decreasing trend.

Indonesia is an archipelago where agricultural industries are spread across several regions in different islands. The changes in meteorological conditions related to climate change issues in each region are different. Analyzing meteorological conditions and crop yields in some crops, i.e., cocoa, coffee, oil palm, cayenne, and paddy, is to study the effect of climate change on agricultural industries. The causal graphs describe that the minimum temperature is possibly affect the production and productivity of cocoa, oil palm, cayenne, and paddy. The minimum temperatures likely influence crop yields of those commodities. Some areas of oil palm producers is also has an annual increasing trend of maximum temperatures. Maximum temperatures have a dependent relationship to oil palm land areas. It indicates that maximum temperatures possibly influence oil palm farming.

In general, long-term changes in meteorological conditions must be faced. The mitigation of climate change in the agricultural sector is an important action to minimize the negative impact of climate change on food security and the economy. Crop failure caused by drought or flood endangers national food security. The most common signs of climate change in some regions of Indonesia include increasing temperatures, decreased wind speeds, drought, and extreme rainfall. Some actions that can be taken encompass developing drought or flood-resistant plant varieties, efficient water management through rainwater harvesting, and the use of organic fertilizers. Therefore, incorporating agriculture and animal husbandry might facilitate the distribution of livestock manure as organic fertilizer. By integrating climate technology, weather forecasting, and sustainable land management, climate disaster mitigation can maintain agricultural productivity while protecting farmers from potential crop failure. The results in this paper are expected to enrich machine learning methods for climate change studies.

The limitation of this paper is that we assume no latent variables in the datasets. It only observes the dependence relationships among variables that exist in the datasets. The main challenge in applying the PC algorithm is that it does not enforce time directionality. The challenges in analyzing time series datasets are observations that are correlated across time and large samples. Another challenge in learning causal graphs from real-world data is that the true underlying graph is unknown. Therefore, to further analyze dependent relationships in the inferred causal graphs, cross-validation is performed using results from related previous studies. In addition, employing y statistical approaches, such as panel regression, helps provide clearer and more robust explanations of the relationships.

Conclusion

In conclusion, the machine learning approach for the integrated analysis of meteorological conditions and agricultural yields has been successfully implemented through the PC algorithm and Intelligent Kernel K-Means for analyzing the impact of climate change on agriculture, a case study in Indonesia. The integrated analysis has been done by observing causal graphs that represent dependence relationships between meteorological variables and crop yield variables for some agricultural commodities, i.e., cocoa, coffee, palm oil, cayenne, and paddy. The main results show that increasing temperatures possibly decrease the yield of cocoa, palm oil, and paddy. This is a warning sign that the mitigation of climate change is needed for sustainable farming. Governments need to develop policies that address rising temperatures in order to support sustainable agriculture. These policies focus on developing heat-resistant crop varieties, improving water management systems such as improving irrigation infrastructure, and adjusting planting schedules. Farmers are encouraged to adopt climate-smart agricultural practices to reduce climate-related risks. Through these measures, agricultural productivity and food security can be maintained despite increasing temperatures. In less vulnerable areas, policies can emphasize capacity building, technology diffusion, and continuous climate monitoring to strengthen national food security in the face of rising temperatures. Future research is encouraged to extend this analytical result to observe the dependence relationships among meteorological data, crop yields, air pollution, and food prices in Indonesia to investigate how climate change affects the broader implications for the national economy.

Data availability

The datasets are available from the corresponding author upon request for strong reasons.

Received: 30 November 2025; Accepted: 12 February 2026
Published online: 16 February 2026

References

1. Hultgren, A. et al. Impacts of climate change on global agriculture accounting for adaptation. *Nature* **642**, 644–652. https://doi.org/10.1038/s41586-025-09085-w (2025).
2. Parolini, G. Weather, climate, and agriculture: Historical contributions and perspectives from agricultural meteorology. *WIREs Clim. Change* **13**, 1–19. https://doi.org/10.1002/wcc.766 (2022).
3. Manesh, M. B. et al. Linkage of agricultural drought with meteorological drought in different climates of Iran. *Theor. Appl. Climatol.* **138**, 1025–1033. https://doi.org/10.1007/s00704-019-02878-w (2019).
4. Houmma, I. H. et al. Temporal relationships between agricultural and meteorological drought over the Oum Er Rbia River, Morrocco. *Big Earth Data* (3), https://doi.org/10.1080/20964471.2025.2479430 (2025).
5. Tran, B.-L., Tseng, W.-C. & Chen, C.-C. Climate change impacts on crop yields across temperature rise thresholds and climate zones. *Sci. Rep.* **15**, 1–17. https://doi.org/10.1038/s41598-025-07405-8 (2025).
6. Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction, and Search* (The MI Press, 2001).
7. Kalisch, M. & Buhlmann, P. Estimating high dimensional directed acyclic graphs with the PC algorithm. *J. Mach. Learn. Res.* **8**, 613–636 (2007).
8. Chickering, D. M. Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3**, 507–554 (2022).
9. Cui, R., Groot, P. & Heskes, T. Copula PC algorithm for causal discovery from mixed data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2016).
10. Handhayani, T. & Cussens, J. Kernel-based approach for learning causal graphs from mixed data. In *Proceedings of the 10th International Conference on Probabilistic Graphical Models, PMLR*, **138**, 221–232 (2020).
11. Liu, W., Tang, M. & Zhao, P. The effects of attention to climate change on carbon, fossil energy and clean energy markets: Based on causal network learning algorithms. *Energy Strateg. Rev.* **59**, 1–13. https://doi.org/10.1016/j.esr.2025.101717 (2025).
12. Yu, H., Jiang, J., Gu, X., Cao, C. & Shen, C. Using dynamic bayesian belief networks to infer the effects of climate change and human activities on changes in regional ecosystem services. *Ecol. Indic.* **170**, 113023. https://doi.org/10.1016/j.ecolind.2024.113023 (2025).
13. Vitolo, C., Scutari, M., Ghalaieny, M., Tucker, A. & Russell, A. Modeling air pollution, climate, and health data using bayesian networks: A case study of the English regions. *Earth Space Sci.* **5**, 76–88. https://doi.org/10.1002/2017EA000326 (2018).
14. Handhayani, T. An integrated analysis of air pollution and meteorological conditions in Jakarta. *Sci. Rep.* https://doi.org/10.1038/s41598-023-32817-9 (2023).
15. Handhayani, T. & Durdane Caglar, K. An analysis of dependence relationships between covid-19 pandemic and environmental conditions cases study in Indonesia, England, and Türkiye. In *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 1–6, 2024, DOI: 10.1109/ISRITI64779.2024.10963452.
16. Sener, E. & Demir, I. Gaussian bayesian network model of healthcare, food and energy sectors in the pandemic: Türkiye case. *Heliyon* **10**, 1–11. https://doi.org/10.1016/j.heliyon.2023.e23798 (2024).
17. Chiang, M.-T. & Mirkin, B. Intelligent choice of the number of clusters in k-means clustering: An experimental study with different cluster spreads. *J. Classif.* **27**, 3–40. https://doi.org/10.1007/s00357-010-9049-5 (2010).
18. Handhayani, T. & Lewenusa, I. An intelligent clustering approach for analyzing a multivariate time series dataset, case study COVID-19 outbreak in Indonesia. In *2023 17th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 1–6, 2023, DOI: 10.1109/TSSA59948.2023.10367007.
19. Handhayani, T., Saputra, A. D. & Jaya, J. Intelligent K-Median clustering for region segmentation of coffee production and consumption in Indonesia. *Procedia Comput. Sci.* **269**, 29–38. https://doi.org/10.1016/j.procs.2025.08.256 (2025).
20. Susilo, S. H., Asrori, A. & Gumono, G. Analysis of the efficiency of using the polycrystalline and amorphous PV module in the territory of Indonesia. *J. Appl. Eng. Sci.* https://doi.org/10.5937/jaes0-31607 (2022).
21. Handhayani, T. & Rusdi, Z. K-means using dynamic time warping for clustering cities in Java island according to meteorological conditions. In *2023 Eighth International Conference on Informatics and Computing (ICIC)*, 1–6, 2023, DOI: 10.1109/ICIC60109.2023.10381899.
22. Handhayani, T. & Lewenusa, I. An analysis of meteorological data in Sumatra and nearby using agglomerative clustering. *J. RESTI (Rekayasa Sist. Dan Teknol. Inf.)* **8**, 234–241. https://doi.org/10.29207/resti.v8i2.5663 (2024).
23. Cordoni, F. & Sancetta, A. Consistent causal inference for high-dimensional time series. *J. Econom.* **246**, 1–18. https://doi.org/10.1016/j.jeconom.2024.105902 (2024).
24. Handhayani, T. et al. A new approach for dynamic analysis of Indonesian food prices using the PC algorithm and vector autoregression. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* **9**, 1417–1427. https://doi.org/10.29207/resti.v9i6.6601 (2025).
25. Handhayani, T., Wasito, I., Sadikin, M. & Ranny. Kernel based integration of gene expression and DNA copy number. In *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 303–308, 2013, DOI: 10.1109/ICACSIS.2013.6761593.
26. Handhayani, T. & Hiryanto, L. Intelligent kernel k-means for clustering gene expression. *Procedia Comput. Sci.* **59**, 171–177. https://doi.org/10.1016/j.procs.2015.07.544 (2015).
27. Handhayani, T. & Wasito, I. Fully Unsupervised Clustering in Nonlinearly Separable Data using Intelligent Kernel K-Means. In *International Conference on Advanced Computer Science and Information System*, 450–453, 2014, https://doi.org/10.1109/ICACSIS.2014.7065891.
28. Shawe-Taylor, J. & Cristianini, N. *Kernel Methods for Pattern Analysis* (Cambridge University Press, 2004).
29. Guo, Y. et al. A panel regression analysis for the COVID-19 epidemic in the United States. *PLoS ONE* **17**, 1–14. https://doi.org/10.1371/journal.pone.0273344 (2022).
30. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H. & Bühlmann, P. Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* **47**, 1–26. https://doi.org/10.18637/jss.v047.i11 (2012).
31. Jordahl, K. et al. geopandas/geopandas: v0.8.1, 2020, DOI: 10.5281/zenodo.3946761.
32. Met Office. Cartopy: a cartographic Python library with a Matplotlib interface. Exeter, Devon (2010–2015).
33. GADM. World / indonesia / sub-divisions (2018).
34. Ashardiono, F. & Trihartono, A. Optimizing the Potential of Indonesian Coffee: A Dual Market Approach. *Cogent Soc. Sci.* **10**, 1–17. https://doi.org/10.1080/23311886.2024.2340206 (2024).
35. Farag, M. A., Mohamed, T. A., El-Hawary, E. A. & Abdelwareth, A. Metabolite Profiling of Premium Civet Luwak Bio-Transformed Coffee Compared with Conventional Coffee Types, as Analyzed Using Chemometric Tools. *Metabolites* **13**, 1–19. https://doi.org/10.3390/metabo13020173 (2023).
36. Jaya, J., Handhayani, T. & Hendryli, J. Mapping Cocoa Harvesting and Export Areas in Indonesia Using Clustering Algorithms. In *International Conference on Information Technology Research and Innovation (ICITRI)*, 1–6, 2025, DOI: 10.1109/ICITRI67507.2025.11232597.
37. Dermoredjo, S. K. et al. The global sway of Indonesian palm oil: An export analysis. *Journal of Agriculture and Food Research* **22**, -, 2025, doi: 10.1016/j.jafr.2025.102064.

38. Tothmihaly, A., Ingram, V. & Taubadel, S. How can the environmental efficiency of Indonesian cocoa farms be increased?. *Ecol. Econ.* **158**, 134–145. https://doi.org/10.1016/j.ecolecon.2019.01.004 (2019).
39. Adejuwon, J. O., Tewogbade, K. E., Oguntoke, O. & Ufoegbune, G. C. Comparing farmers' perception of climate effect on cocoa yield with climate data in the humid zone of Nigeria. *Heliyon* **9**, 1–15. https://doi.org/10.1016/j.heliyon.2023.e23155 (2023).
40. Lander, T. A. et al. Global chocolate supply is limited by low pollination and high temperatures. *Communications Earth & Environment* **6**, 1–8. https://doi.org/10.1038/s43247-025-02072-z (2025).
41. Megerssa, G. R., Garedew, W., Jespersen, K., Grabs, J. & Bekele, A. E. Impact of climate change on Ethiopian Arabica coffee production and current challenges it poses to coffee value chain. *Sustainable Futures* **10**, 1–17. https://doi.org/10.1016/j.sftr.2025.101459 (2025).
42. Paterson, R. R. M. & Lima, N. Climate change affecting oil palm agronomy, and oil palm cultivation increasing climate change, require amelioration. *Ecol. Evol.* **8**, 452–461. https://doi.org/10.1002/ece3.3610 (2018).
43. Nugroho, B. D. A. et al. Effect of climate variability on Chili Pepper (*Capsicum frutescens* L.) cultivation in Indonesia. *Environ. Nat. Resour. J.* **23**, 459–468 (2025).
44. Puspitaningrum, D., Reka, W. Y. & Widayanto, B. Impact of climate change on cayenne pepper availability in sleman district, Yogyakarta with dynamic system modelling. In *7th International Conference on Agriculture, Environment, and Food Security*, **1302**, 1–7, 2024, DOI: 10.1088/1755-1315/1302/1/012131.
45. Krishnan, P., Ramakrishnan, B., Reddy, K. R. & Reddy, V. Chapter three - High-Temperature Effects on Rice Growth, Yield, and Grain Quality. *Advances in Agronomy* **111**, 87–206. https://doi.org/10.1016/B978-0-12-387689-8.00004-7 (2011).
46. Yuliawan, T. & Handoko, I. The effect of temperature rise to rice crop yield in Indonesia uses Shierary Rice Model with Geographical Information System (GIS) feature. *Procedia Environ. Sci.* **33**, 214–220. https://doi.org/10.1016/j.proenv.2016.03.072 (2016).
47. Oh, D. et al. Effect of elevated air temperature on the growth and yield of paddy rice. *Agronomy* **13**, 1–14. https://doi.org/10.3390/agronomy13122887 (2023).

Acknowledgements

This research is supported by a Fundamental Research Grant 2025 decree No 0419/C3/DT/.05.00/2025 and agreement/contract No 124/C3/DT.05.00/PL/2025;1011/LL3/AL.04/2025;0584/Int-KLPPM/UNTAR/VI/2025 from the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia.

Author contributions

T. Handhayani works on data collection, experiments, analysis, and writing manuscripts. D. Arisandi contributed to the analysis and writing manuscript. Wasino Wasino contributed to the analysis.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-40418-5.

Correspondence and requests for materials should be addressed to T.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026