AgriEngineering

Article

A Data-Driven Approach to Improve Cocoa Crop Establishment in Colombia: Insights and Agricultural Practice Recommendations from an Ensemble Machine Learning Model

Leonardo Talero-Sarmiento [1,*], Sebastian Roa-Prada [2], Luz Caicedo-Chacon [3] and Oscar Gavanzo-Cardenas [4]

[1] Facultad de Ingenieria, Ingenieria Industrial, Universidad Autonoma de Bucaramanga (UNAB), Bucaramanga 680003, Colombia
[2] Facultad de Ingenieria, Ingenieria Mecatrónica, Universidad Autonoma de Bucaramanga (UNAB), Bucaramanga 680003, Colombia; sroa@unab.edu.co
[3] Ingenieria de Sistemas, Fundacion Universitaria de San Gil (Unisangil), San Gil 684031, Colombia; lcaicedo@unisangil.edu.co
[4] FEDECACAO, Bucaramanga 680006, Colombia; formacion_cacao1@fedecacao.com.co
* Correspondence: ltalero@unab.edu.co; Tel.: +57-607-643-6111 (ext. 309)

Abstract: This study addresses the critical challenge of the limited understanding of environmental factors influencing cocoa cultivation in Colombia, a region with significant production potential but diverse agroecological conditions. The fragmented nature of the existing agricultural data and the lack of targeted research hinder efforts to optimize productivity and sustainability. To bridge this gap, this research employs a data-driven approach, using advanced machine learning techniques such as supervised, unsupervised, and ensemble models, to analyze environmental datasets and provide actionable recommendations. By integrating data from official Colombian sources, as well as the NASA POWER database, and geographical APIs, the present study proposes a methodology to systematically assess environmental conditions and classify regions for optimal cocoa cultivation. The use of an assembled model, combining clustering with targeted machine learning for each cluster, offers a more precise and scalable understanding of cocoa establishment under diverse conditions. Despite challenges such as limited dataset resolution and localized climate variability, this research provides valuable insights for a more comprehensive understanding of the environmental conditions impacting cocoa plantation establishment in a given location. The key findings reveal that temperature, humidity, and wind speed are crucial determinants of cocoa growth, with complex interactions affecting regional suitability. The results offer valuable guidance for the implementation of adaptive agricultural practices and resilience strategies, enabling sustainable cocoa production systems. By implementing better practices, countries such as Colombia can achieve higher market shares under growing global cocoa demand conditions.

Keywords: cocoa production; Colombia; climate change; data-driven agriculture; ensemble model; machine learning; environmental impact; sustainable farming practices

Academic Editors: Muhammad Jehanzeb Masud Cheema, Muhammad Aqib, Ahmed Elbeltagi, Shoaib Rashid Saleem and Saddam Hussain

Received: 31 October 2024
Revised: 21 December 2024
Accepted: 26 December 2024
Published: 28 December 2024

Citation: Talero-Sarmiento, L.; Roa-Prada, S.; Caicedo-Chacon, L.; Gavanzo-Cardenas, O. A Data-Driven Approach to Improve Cocoa Crop Establishment in Colombia: Insights and Agricultural Practice Recommendations from an Ensemble Machine Learning Model. *AgriEngineering* **2025**, *7*, 6. https://doi.org/10.3390/agriengineering7010006

1. Introduction

Cocoa holds substantial economic and cultural significance globally, primarily as the core ingredient in chocolate production, a commodity experiencing increasing demand worldwide. Côte d'Ivoire and Ghana dominate global production, collectively contributing over 70% of the world's cocoa supply [1]. In Latin America, Brazil and Ecuador are prominent producers [2]. Colombia has emerged as a growing player, with cocoa cultivation

occurring in 29 of its 32 states. Santander, Arauca, Antioquia, Huila, Nariño, and Tolima are key production regions, with Santander alone accounting for 34.4% of the national output in 2023 [3].

Renowned for its fine-flavored cocoa beans, characterized by fruity, bitter, and acidic notes [4], Colombia enjoys a prestigious position in the international market [5]. The superior quality of Colombian cocoa beans stems from a combination of genetics, climate, and soil conditions, which create optimal growing environments in specific geographic areas [6]. Despite these advantages, most of Colombia's exports fall within the category of ordinary cocoa [7]. Previous research has highlighted productivity and quality as critical bottlenecks in the Colombian cocoa agri-food value chain [8]. Government initiatives promoting sustainable agricultural practices have gradually improved yield and bean quality, allowing Colombian cocoa to expand its global market share. However, the industry continues to grapple with significant challenges, including fluctuating market prices [9], limited access to financial resources for smallholder farmers [10], and inadequate infrastructure in rural areas [11]. Compounding these issues is the threat posed by climate change, which introduces greater variability and exacerbates the existing vulnerabilities in agricultural systems [12]. Climate change, characterized by an increased frequency of droughts, floods, and storms, poses significant risks to agriculture both in Colombia and globally [13].

In the face of climate change, cocoa farmers encounter a multitude of challenges that reduce yields, such as higher pest and disease incidence [14–16], aging farms and trees [17], and the use of low-yield planting materials [18,19]. Diminishing soil fertility due to poor nutrition [20,21] and inadequate plantation densities [22,23] further exacerbate these issues. Cocoa cultivation, which thrives under stable temperatures and adequate rainfall, is particularly vulnerable to climate fluctuations [24–26]. Projections suggest that in 2050, suitable cocoa-growing areas could shrink in regions such as the Amazon, necessitating crop adaptation or changes in agronomic practices [27,28].

Innovative strategies are being explored to mitigate these effects, such as developing climate-resilient cocoa varieties [29], improving soil management within agroforestry systems [30], and optimizing irrigation practices [31]. Despite these efforts, the need for data-driven approaches remains critical, especially in identifying which environmental variables most strongly impact cocoa production across different regions [32,33]. While the existing studies acknowledge cocoa's vulnerability to climate change, a notable gap persists in understanding the specific environmental factors influencing yields in Colombia's diverse agroecological zones [34].

To address these challenges, advanced technologies, including blockchain, IoT, Big Data, and Machine Learning (ML), have been introduced to agricultural practices [35,36]. Machine Learning is a promising approach to improving cocoa production, offering tools for yield prediction, disease detection, and efficient resource management [37]. However, implementing ML in cocoa farming presents several challenges. A primary obstacle is the lack of comprehensive datasets for training ML models, especially compared to staple crops like wheat or maize, for which ample data sources are available [38]. The variability in environmental conditions across different cocoa-growing regions also complicates the creation of generalized models [39]. Additionally, smallholder farmers face limitations in accessing advanced technologies and real-time data, which are essential for making informed decisions. Bridging these gaps with intuitive, localized tools and infrastructure investments is vital for increasing cocoa productivity and ensuring sustainability.

One promising approach for enhancing the performance of machine learning models, while reducing computational complexity, is assembling models [40]. Assembled models, which combine multiple machine learning algorithms, such as bagging, boosting, and stacking, have significantly improved prediction accuracy and robustness in various

fields. These techniques are particularly advantageous in addressing crop production challenges, where integrating multiple weak learners can help improve the reliability of yield predictions and environmental impact assessments [41].

Despite their potential, the implementation of assembled models for the analysis of cocoa production establishment remains limited. The complexity of the data architecture [42], the fragmented nature of agricultural information generating data scarcity [43], and the lack of sufficient computational resources are major challenges to adopting these advanced ML approaches [44]. Furthermore, most research in cocoa farming—especially in production activities—has focused on simpler, standalone models that are easier to interpret and require less computational resources [45–48].

The limited use of assembled models represents an important gap in the current research landscape. Addressing this gap could significantly increase model accuracy and the reliability of cocoa models for estimating yield, quality, and resilience, especially regarding climate variability. By supporting more advanced computational methods and encouraging collaborations between agronomists and data scientists, the cocoa sector can begin to harness the full potential of assembled ML models to tackle the complex challenges it faces.

Thus, assembled models also offer significant potential for defining timely interventions before cocoa crop establishment. The current research highlights several agronomic practices aimed at improving cocoa productivity, such as optimizing shade management through agroforestry [49,50], enhancing soil moisture retention through mulching [51], and integrating pest management to reduce crop losses [52]. These practices help create favorable plantation microclimates for optimal flower and fruit development under variable environmental conditions [50,53]. However, there is still a gap in determining the adaptability and effectiveness of these practices across different cocoa-growing regions, particularly those with challenging environmental conditions.

For instance, while shade management has proven beneficial, determining the optimal levels and types of shading that maximize yields under extreme conditions, such as high solar irradiance and low humidity, still requires further investigation. By analyzing complex datasets from multiple regions, assembled models can offer insights into the most effective interventions tailored to specific environmental contexts. Leveraging these models can lead to more precise, data-driven recommendations for farmers, improving overall crop establishment success and productivity.

This research aims at evaluating environmental data and assessing its impact on cocoa production using advanced analytics and machine learning, including assembled models. Open-access datasets, such as meteorological data, agricultural outputs, and soil suitability indicators sourced from the NASA POWER database, can classify Colombian regions based on their suitability for cocoa cultivation. Techniques like logistic regression, decision trees, random forests, support vector machines (SVMs), and neural networks are integrated into assembled models to identify key climatic factors affecting yield.

2. Materials and Methods

This study focuses on the geographical suitability of cocoa cultivation across Colombia, specifically targeting the identification and analysis of optimal growing areas. The study area encompasses various regions across Colombia, characterized by diverse climates, soil types, and topographical conditions. The climate varies from humid tropical to dry regions, influencing the suitability for cocoa growth. Soil types include well-drained loamy soils, generally favorable for cocoa, and areas with clay and sandy soils.

Data were systematically acquired through the official Colombian Open-Data program using a dedicated Application Programming Interface (API) [54]. The primary dataset

consisted of several variables, including spatial geometry in multi-polygon format, administrative identifiers, geographic codes, land area measurements, and land suitability classifications for cocoa cultivation. Specific sites included various municipalities across cocoa-producing regions, such as Antioquia, Santander, and Huila, representing diverse environmental conditions.

Centroid points were selected as representative markers of each region for data representation. The selection depended on the characteristics of each polygon, particularly its convexity and compactness. The centroid provided a precise representation in convex polygons (in which all interior angles are less than 180°). Furthermore, compactness metrics and additional geometric evaluations, such as the moment of inertia and the compaction index, were also employed to accurately represent the topographical conditions for more complex, non-convex polygons. The maximum inner circle was calculated for highly irregular polygons, and its center was chosen as the representative point. Figure 1 shows four illustrative examples, while Figure 2 presents the overall process for selecting the representative points. This study used a compactness threshold of 0.5, and compactness values were calculated for polygons with low balance.

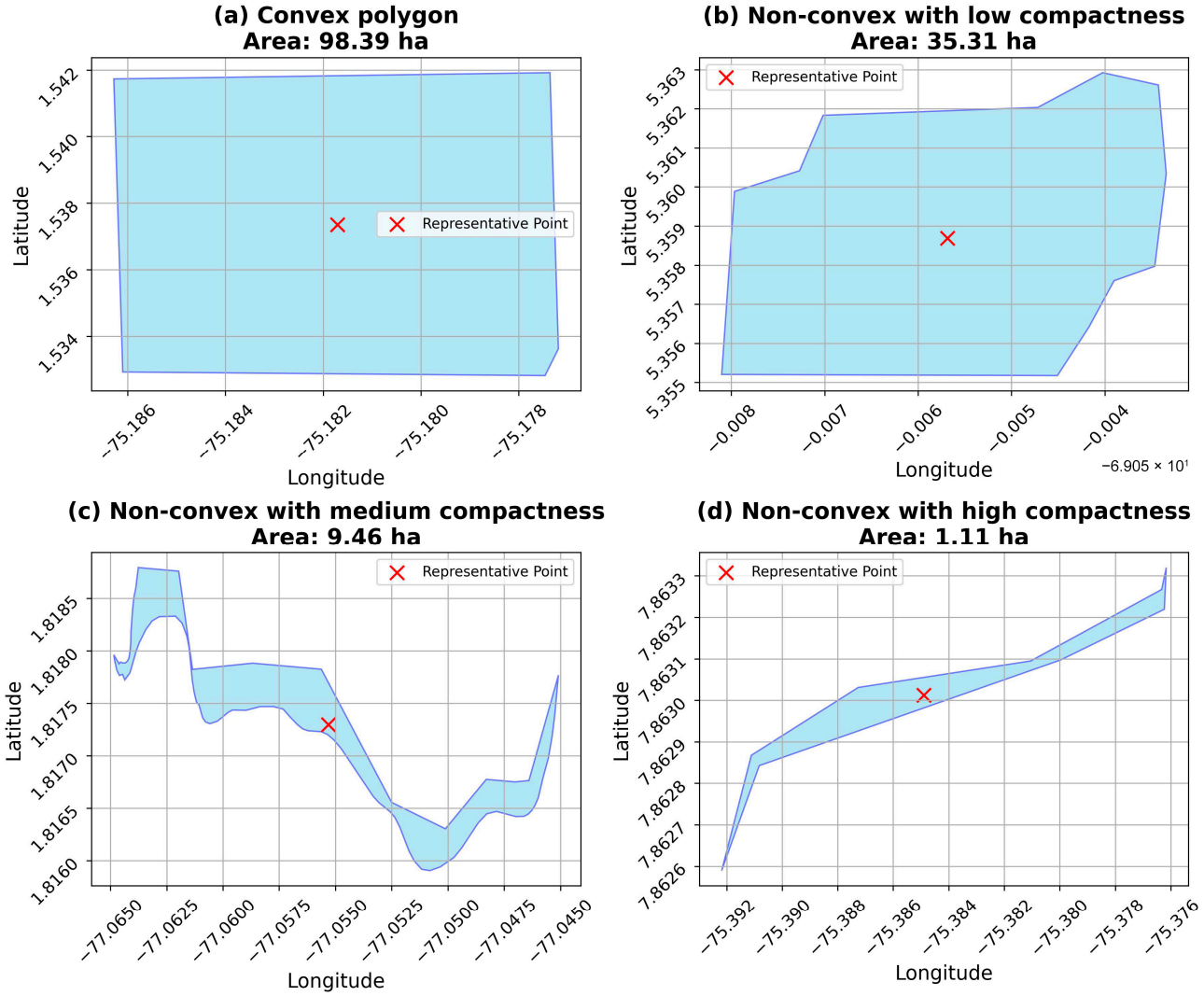

Figure 1. Examples of polygon shapes and compactness for potential cocoa cultivation areas.

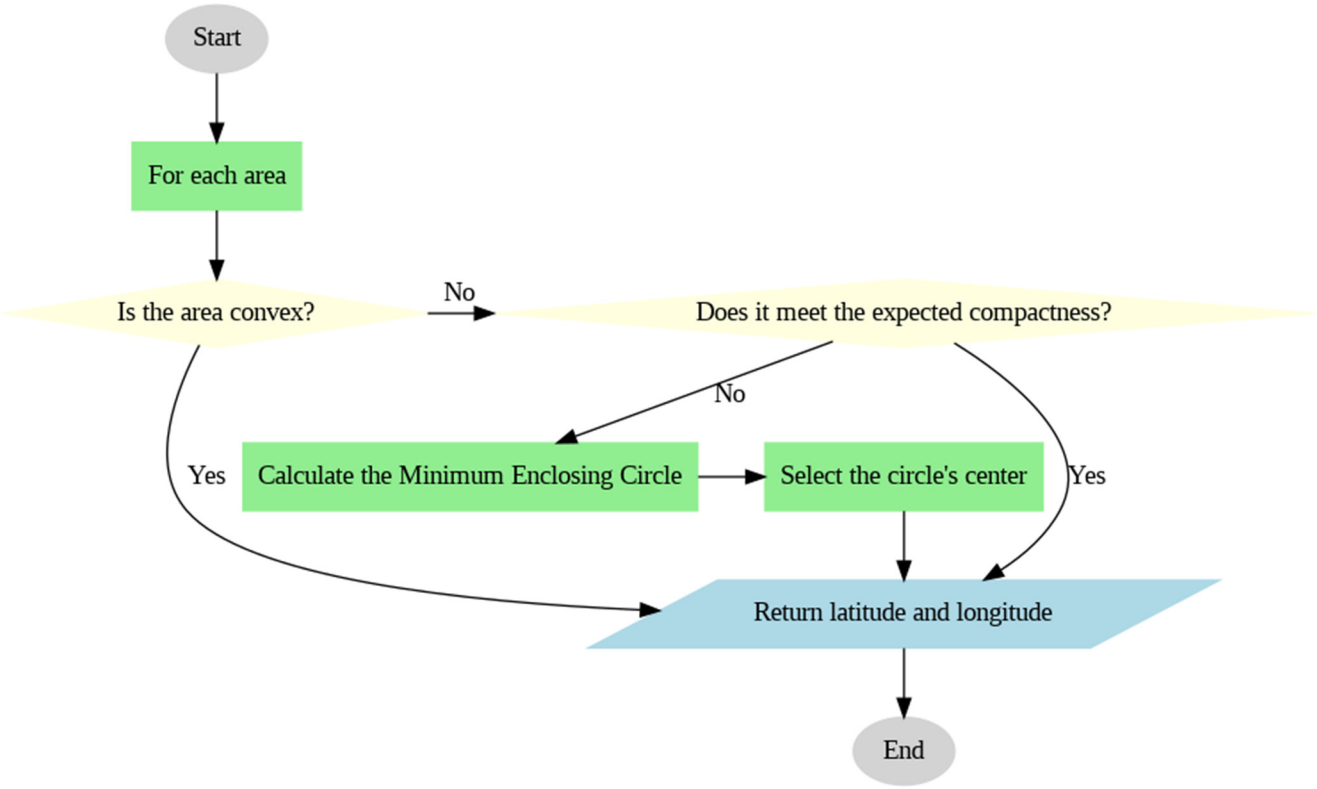

Figure 2. Flowchart for the representative point selection process.

2.1. Data Collection

The data collection focused on gathering essential variables related to cocoa cultivation, including meteorological, soil, and crop-related data. Meteorological data were retrieved from the NASA POWER database, providing variables such as solar radiation, precipitation, temperature, relative humidity, and wind speed. The NASA POWER database offers meteorological data with a spatial resolution of 0.5° by 0.5°, approximately 55 km by 55 km at the equator. While suitable for regional analyses, this resolution may not capture local microclimatic variations, which are particularly relevant in cases of precision agriculture and crops sensitive to treatment applications [55].

Soil data included soil moisture content, surface soil wetness, and root zone wetness, which were also obtained from the NASA dataset at a spatial resolution of 0.25° by 0.25° (approximately 27.75 km by 27.75 km at the equator). Additionally, elevation data were integrated for each representative point using the Open-Elevation API to accurately measure altitude, which plays a critical role in cocoa cultivation. The data collection from 2019 to 2023 focuses on the most recent and complete data without missing values to ensure accuracy. The instruments used included APIs for data retrieval, specifically designed to provide high-resolution spatial and environmental information. The variables analyzed with their respective codification include the following:

- ALLSKY_SFC_SW_DWN (All-Sky Surface Shortwave Downward Irradiance): Measures the solar radiation reaching the Earth's surface (kWh/m^2/day), which is crucial for photosynthesis in cocoa plants [56].
- PRECTOTCORR (Corrected Total Precipitation): Quantifies precipitation (mm), which impacts irrigation and natural soil moisture levels, essential for cocoa's well-watered growth conditions [57].

- CLRSKY_SFC_PAR_TOT (Clear-Sky Photosynthetically Active Radiation): Estimates light availability (W/m^2) for photosynthesis, indicating the potential for cocoa growth in optimal sunlight [58].
- RH2M (2 m Relative Humidity): Represents atmospheric moisture content (%), influencing transpiration and pest/disease development. It is critical for cocoa's high-humidity needs [59].
- WS2M (2 m Wind Speed): Measures wind speed (m/s), which affects evapotranspiration and microclimate conditions. It is essential for pollination and fungal disease prevention [60].
- T2M_MAX and T2M_MIN (Maximum and Minimum Temperature at 2 Meters): Tracks temperature (°C), as stable temperatures are vital for cocoa growth and fruit development [61].
- GWETTOP (Surface Soil Wetness): Assesses soil moisture in the top 5 cm, indicating water availability for cocoa seedlings [61].
- GWETPROF (Root Zone Soil Wetness): This variable refers to the soil moisture content, spanning from the surface to a depth of 100 cm. This measurement encompasses the primary root zone where most mature cocoa plant roots are located, making it pivotal for assessing water availability to support healthy plant growth [61].
- GWETROOT (Profile Soil Moisture): Measures total soil moisture from the surface to bedrock, providing a long-term view of water supply [61].

This dataset, comprising 57,658 records, was meticulously curated to identify suitable zones for cocoa production in Colombia. The dependent variable, "aptitude", categorized land into high, medium, and low suitability levels. Additional features, including altitude, soil characteristics, and twelve environmental variables, were detailed with their mean and standard deviation values across the dataset's 81 columns, capturing the complexity of factors influencing cocoa production.

2.2. Experimental Design

The experimental design featured two main parts. The first part involved selecting the best ML model to estimate the suitability (aptitude) for establishing cocoa crops by comparing performance on balanced and unbalanced datasets. A battery of ML models, including logistic regression, decision trees, random forests, SVMs, and neural networks, was applied. The second part of the design involved proposing an assembled model. Clustering was performed to group similar regions within Colombia based on environmental and geographical features. After clustering, the same battery of ML models was applied to classify cocoa suitability within each cluster, followed by training a separate model to classify the regions into these clusters based on location characteristics. The assembled model consisted of two stages: (1) classification of locations into defined clusters and (2) aptitude classification for each location within its corresponding cluster.

The dataset was divided into training and testing subsets using a 10-fold cross-validation method to validate model accuracy and reliability. Control variables included climatic and soil factors that were kept constant, while the primary variation was geographic location and corresponding environmental conditions. Advanced machine learning classification techniques were employed to analyze this dataset, including the following:

- Logistic Regression (LR): This linear classifier utilized the logistic function to predict the probability of the "aptitude" categories [62]. The logistic regression model was implemented with standardized regression coefficients, providing insight into the relative importance of each feature. To ensure convergence, the model was configured to run with a maximum of 10,000 iterations, making it suitable for datasets with high dimensionality like ours.

- Decision Tree Classifier: The decision tree (DT) technique implemented had a maximum depth of 20 levels, allowing it to capture complex patterns in the data while avoiding overfitting. The tree's splitting criterion was based on the Gini impurity index, which measures the purity of the node's splits. Due to its inherent feature importance metric, this model is particularly useful for interpreting which variables significantly impact predicting land suitability [63].
- Random Forest Classifier: The random forest (RF) model, an ensemble of multiple decision trees, was configured with a maximum depth of 15 per tree for the aptitude problems and a maximum depth of 150 for the cluster classification problem. It used the same Gini impurity criterion for node splitting as the decision tree model. The ensemble approach of random forests, where multiple trees vote for the most popular class, enhances the model's robustness and reduces the risk of overfitting. This model was instrumental in identifying key variables through aggregate feature importance scores across the trees [64].
- Support Vector Machine (SVM): The SVM technique was configured with a radial basis function (RBF) kernel, which is well-suited for handling non-linear relationships within the data. This kernel maps the input features into higher-dimensional space, allowing the SVM to find the optimal hyperplane that maximizes the margin between classes. The SVM was also standardized to ensure that the features contributed equally to the decision boundary [65].
- Neural Network (MLPClassifier): The Artificial Neural Network (ANN) model employed a Multi-Layer Perceptron (MLP) architecture [66] with one hidden layer and a ReLU (Rectified Linear Unit) activation function. The ReLU function, known for mitigating the vanishing gradient problem, allowed the model to learn complex patterns within the data. The model was trained with a maximum of 10,000 iterations, ensuring thorough training and convergence.
- The k-means clustering algorithm was utilized to segment the data into distinct clusters based on climate and soil characteristics [67]. The preprocessing pipeline included standardizing features and regularizing the covariance matrix by means of the LedoitWolf estimator method, ensuring numerical stability [68]. The data were divided into sub-dataframes of 10,000 records to handle the large dataset efficiently. The elbow method and silhouette score determined the optimal number of clusters, ranging from 2 to 100, by calculating average inertia and silhouette scores across batches. After selecting the optimal number of clusters based on silhouette scores, final k-means clustering was conducted, allowing us to identify the inherent data structures.

2.3. Data Analysis

Exploratory data analysis (EDA) was conducted to uncover patterns, distributions, and key features in the dataset. It involved visualizing environmental variables, such as temperature, precipitation, and solar radiation, to understand their influence on cocoa cultivation suitability. Heatmaps, boxplots, and confidence interval plots were used to identify correlations between variables and detect anomalies. Multivariate data analysis was performed using clustering techniques to group similar regions based on environmental and geographical characteristics, excluding latitude and longitude. This approach allowed for identifying distinct regions with comparable conditions, which is essential for refining the classification of cocoa suitability. Data analysis was conducted using Python, incorporating libraries like scikit-learn for machine learning, pandas for data manipulation, and Matplotlib for visualization [69–71]. Statistical methods included descriptive analysis (mean, standard deviation) and predictive modeling through machine learning models.

2.4. Assembled Model Description

The assembled model was developed in three stages to improve the classification of cocoa cultivation suitability across different regions in Colombia.

2.4.1. Stage 1: Cluster Analysis

Initially, clustering analysis was conducted to define groups of regions with similar environmental and geographical characteristics. This clustering step allowed for a more tailored analysis of regions, helping capture nuanced climate, soil, and topography differences that influence cocoa suitability. The clustering process utilized k-means to group locations, enabling the subsequent modeling steps to leverage this categorization for refined analysis.

2.4.2. Stage 2: Cluster Classification

After the cluster definition step, several machine learning models were trained to classify each location into the appropriate cluster based on its geographical and environmental attributes. The model with the best performance acted as the main model for cluster classification, and the remaining four models were employed as predictors, so the cluster predictions from these models were then integrated back into the original dataset as additional features, thereby enhancing the dataset with cluster-specific information that enriched the subsequent analysis.

2.4.3. Stage 3: Machine Learning Model Training per Cluster

The enriched dataset—including original features and cluster assignments—was used in the third stage to train machine learning models specific to each cluster. By training separate models for each cluster, the unique characteristics of each group were considered, resulting in improved model performance. A random forest model with a maximum depth of 150 was employed to classify cocoa suitability within each cluster. Machine learning models were also applied to each cluster to refine the classification. Training and testing models within each cluster aimed at accurately capturing local variations and improving the precision of cocoa suitability predictions. Figure 3 shows a representation of the proposed assembled model.

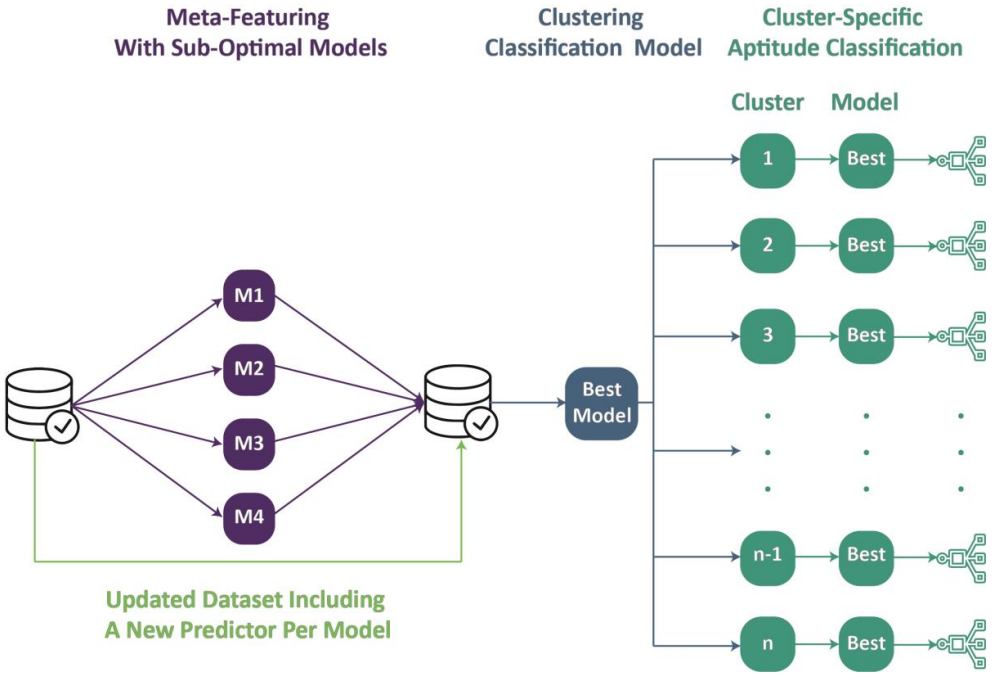

Figure 3. Proposed assembled model architecture.

2.5. Robustness Analysis Using K-Folds

Robustness analysis was conducted using k-fold cross-validation to evaluate the stability, consistency, and reliability of the machine learning models applied in this study. A 10-fold cross-validation approach was employed, involving the division of the dataset into ten equal subsets. Each model was trained on nine of these subsets and tested on the remaining one, with this process repeated ten times to ensure each subset served as a test set once. This approach not only helped in assessing the consistency of the models but also provided a comprehensive measure of performance. Multiple metrics were employed during k-fold cross-validation to understand model robustness thoroughly. These included the following:

- Accuracy: Measured the proportion of correctly predicted instances out of the total instances, providing a general measure of how well the model performed across all classes.
- Precision: Calculated as the ratio of true positives to the sum of true positives and false positives, indicating how many predicted positive cases were correct.
- Recall: Also known as sensitivity or the true positive rate. Measured the ratio of correctly predicted positive observations to all actual positives, reflecting the model's ability to identify positive cases.
- F1-Score: Represented the harmonic mean of precision and recall, balancing these two metrics and providing a single measure of a model's predictive performance, which is especially useful when dealing with imbalanced datasets.
- Cross-Validation Scores: Cross-validation accuracy was averaged across the ten folds to estimate model generalizability. Using cross-validation scores was crucial to determine the variance and ensure that the model did not overfit to a particular subset of the data.

Figure 4 summarizes and organizes the process from data acquisition to agricultural practice recommendations, integrating the ensemble machine learning model. The light gray color represents the main activities; light blue indicates the subprocesses in each activity, and light green denotes the decisions.

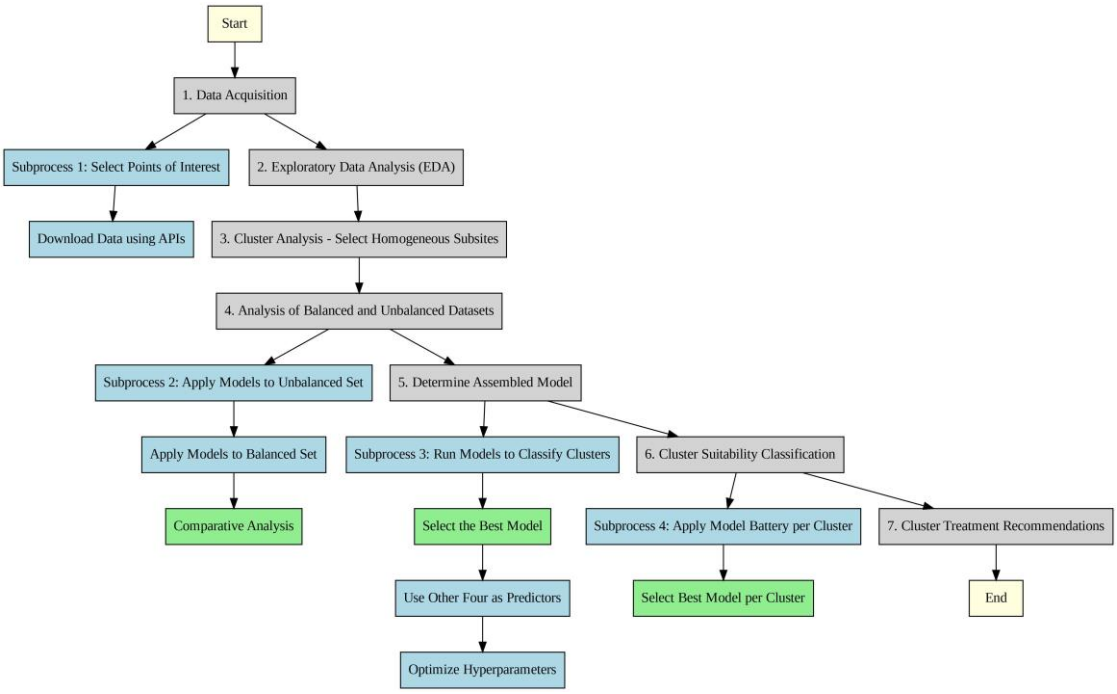

Figure 4. Research methodology flowchart.

2.6. Model Performance Evaluation

The machine learning experiments for classifying the aptitude of a terrain for cocoa farming in Colombia were conducted using the cloud-based platform Google Colaboratory. The virtual machine provided for the computations was configured with a Linux-based operating system, specifically Linux-6.1.85+-x86_64-with-glibc2.35, ensuring a robust and secure environment for data processing. The analysis was performed using Python 3.10.12, leveraging its ecosystem of libraries for data manipulation and machine learning. The computational core was powered by an Intel(R) Xeon(R) CPU @ 2.20GHz, equipped with 12.67 GB of RAM, which facilitated efficient handling of the environmental datasets. The available disk space on the virtual machine was 107.72 GB, ample for the dataset and intermediate outputs.

The specifications were adequate to handle the datasets and computations without significant latency. The reliance on CPU-based computations rather than GPU acceleration underscored the efficiency of the chosen models and preprocessing steps. For example, the computational setup enabled iterative model training and validation within a reasonable timeframe of approximately two hours for the entire suite of experiments. This demonstrates that the system could handle the inherent complexity of environmental datasets and multiple modeling techniques, including random forest and neural networks, which are computationally intensive. Despite lacking a GPU, the CPU and memory resources allowed the models to effectively process both balanced and unbalanced datasets. Table 1 explains the parameter selection for each model.

Table 1. Model parameters and their relevance in the classification of cocoa aptitude.

Model	Parameters	Default Settings Not Mentioned	Relevance to the Classification Problem
Decision Tree	max_depth = 20, criterion = gini, splitter = best	min_samples_split = 2, min_samples_leaf = 1	Increased depth helps manage many predictors by allowing more complex decision paths.
Random Forest	max_depth = 15, criterion = gini, n_estimators = 100	min_samples_split = 2, min_samples_leaf = 1	Depth and multiple estimators improve accuracy and prevent overfitting with many correlated variables.
Neural Network	hidden_layer_sizes = (100,), activation = relu, solver = adam	learning_rate_init = 0.001, max_iter = 10,000	Layers and neurons can model complex non-linear relationships in high-dimensional data.
Logistic Regression	penalty = l2, C = 1.0, solver = lbfgs	max_iter = 10,000, multi_class = auto	Regularization (L2) and iterative solver handle multicollinearity, enhancing stability.
Support Vector Machine	kernel = rbf, C = 1.0	gamma = scale	The RBF kernel efficiently manages high-dimensional space, ideal for non-linear data separation.

3. Results

3.1. Exploratory Data Analysis

The land suitability analysis for cocoa cultivation depends on the geographic location of different study areas. Figure 5 illustrates the spatial distribution of potential cocoa cultivation zones across Colombia, representing the farms selected by grey points. Figure 6 shows significant trends in the distribution of potential cocoa cultivation areas, emphasizing the role of smallholder farms. The data reveal that areas equal to 5 hectares or less correspond to the lower 20% available land, aligning with the global trend of smallholder farmers dominating cocoa production. This indicates that much of Colombia's cocoa-growing potential lies in smaller regions, likely integral to the local cocoa economy. However, as the percentile increases, the available land size grows rapidly, with areas exceeding 46 hectares by the 50th percentile and over 100 hectares by the 75th percentile, indicating land suitable for agribusiness. Despite the availability of these larger areas, their use for cocoa cultivation may be limited. Additionally, the data show a steady increase in

altitude, with most large land tracts situated above 800 m above sea level (a. s. l.), some even exceeding 1400 m a. s. l. While cocoa can be grown at these altitudes, it may require specialized varieties or advanced agricultural practices.

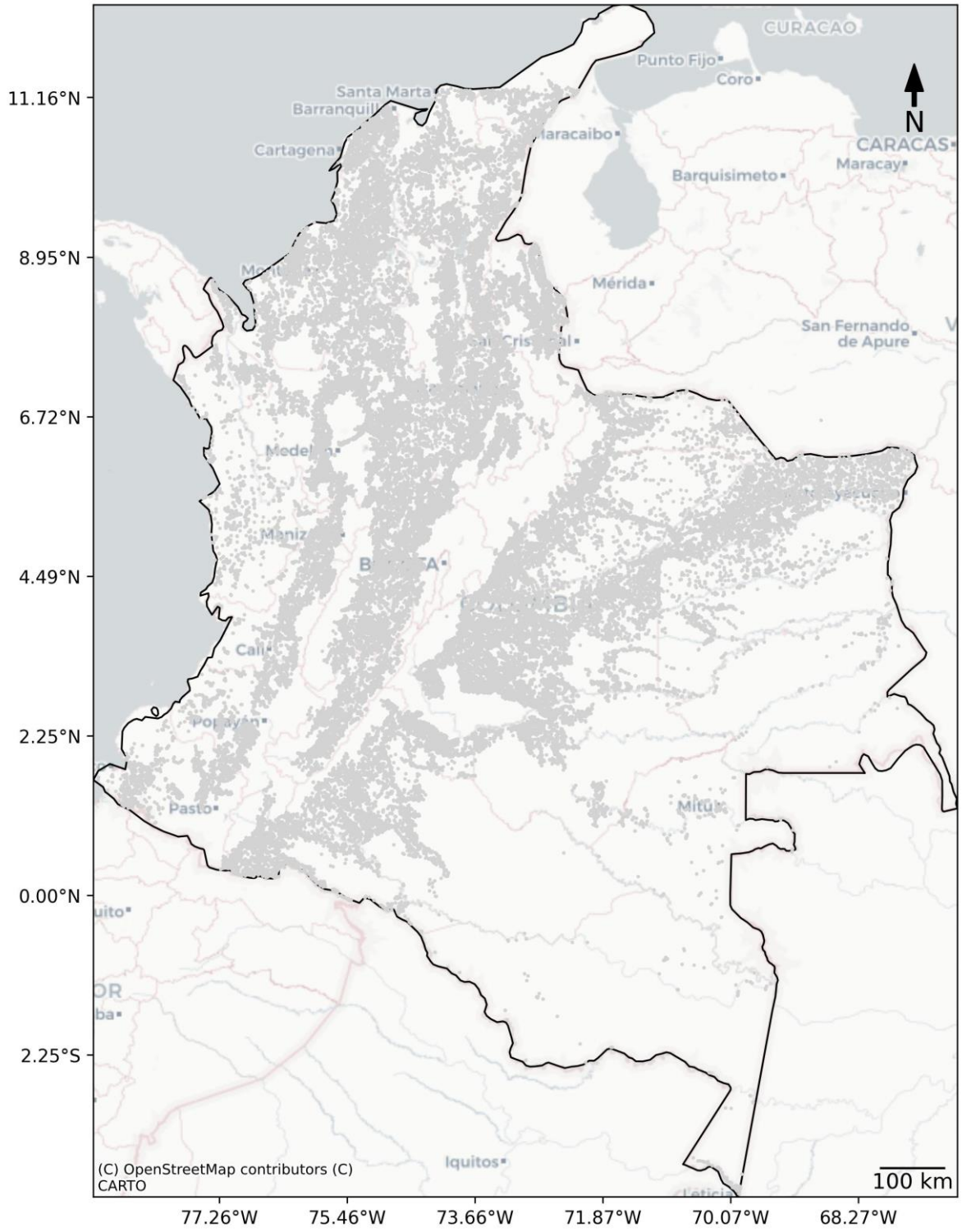

Figure 5. Geographic distribution of potential cocoa cultivation areas in Colombia.

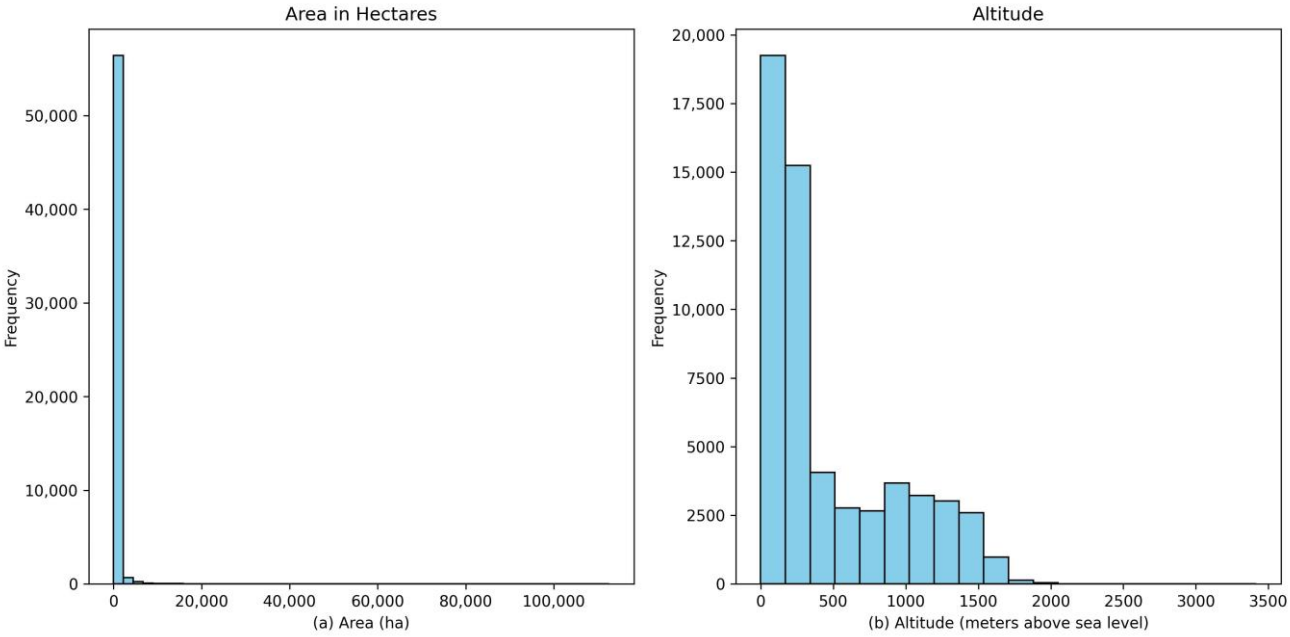

Figure 6. Distribution of potential cocoa cultivation areas by size and altitude in Colombia.

3.1.1. Climatological Analysis

The climatological data provide critical insights into the relationship between environmental factors and cocoa cultivation. The analysis focuses on how wind speed, temperature, relative humidity, precipitation, and solar radiation yearly variations could impact cocoa crop establishment and its future production in different areas of Colombia (Figure 7 contains the average and confidence interval, while Figure 8 shows a yearly boxplot for each environmental variable).

- Wind Speed and Pollination: As shown in Figure 7, variability in wind speed significantly affects cocoa pollination. Cocoa relies on wind and insects for pollination, and erratic wind patterns can disrupt these processes. High wind speeds can damage flowers and young pods, reducing yields and increasing water loss through increased leaf transpiration.
- Temperature and Evapotranspiration: The data suggest that increasing temperature variability, particularly extremes, can elevate evapotranspiration (ET_0) rates, leading to higher water demand for cocoa plants. If the water supply does not meet these demands, it can result in plant stress, reduced photosynthetic efficiency, and yield losses. Additionally, temperature extremes can accelerate or delay cocoa development, disrupting the phenological cycle and yield quality.
- Relative Humidity and Disease Prevalence: Fluctuations in humidity levels can create conditions favorable for diseases like witches' broom and frosty pod rot, which thrive in high humidity conditions. Periods of elevated humidity, especially when coupled with high temperatures, can exacerbate these diseases, posing a significant risk to cocoa production.
- Precipitation and Solar Radiation: Decreasing precipitation and solar radiation trends could potentially challenge cocoa establishment. Insufficient rainfall may lead to inadequate soil moisture, affecting water availability and increasing plant stress. Meanwhile, lower solar radiation reduces the energy available for photosynthesis, which is essential for plant growth and productivity.

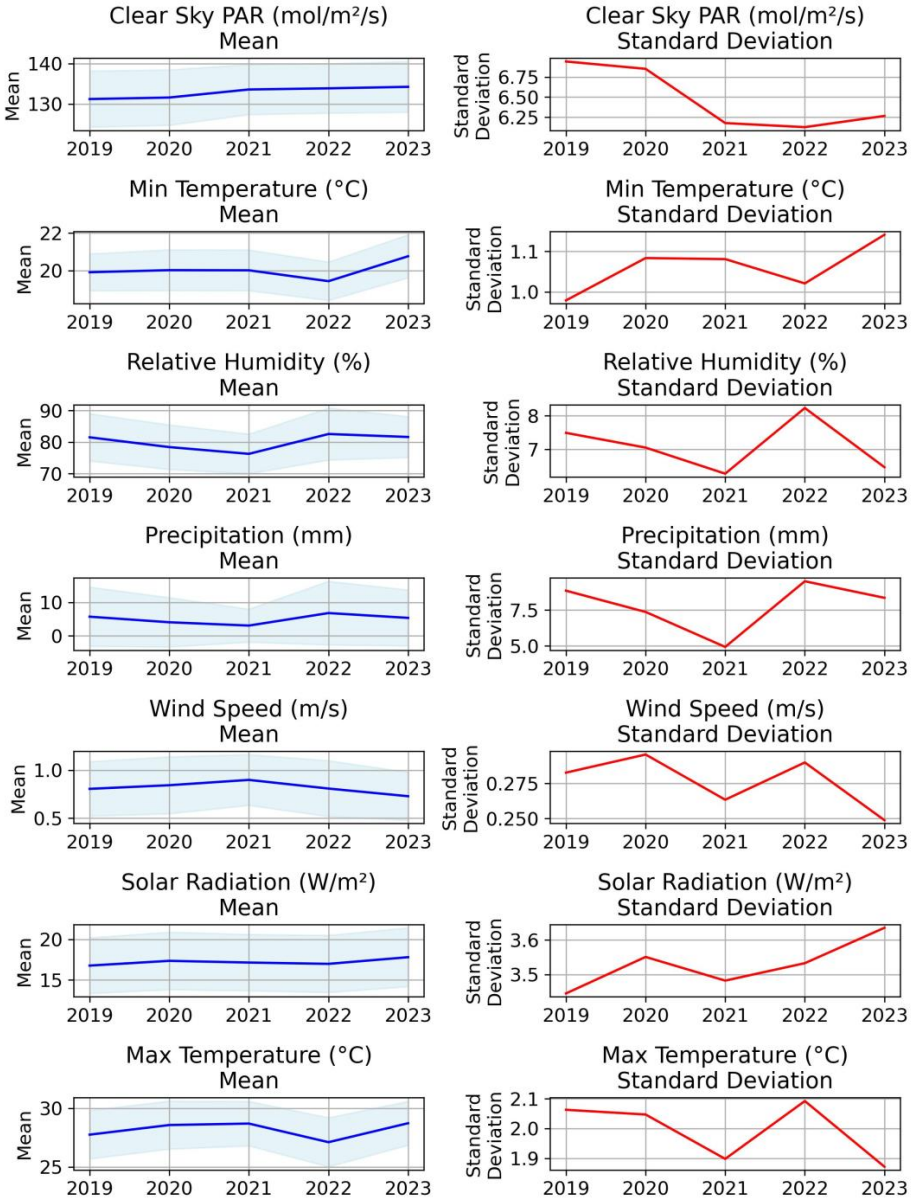

Figure 7. Temporal variability in key climatic variables affecting cocoa cultivation (2019–2023).

Decisionmakers must contemplate environmental characteristics to guarantee optimal establishment. Soil fertility and quality are pivotal for optimal cocoa production, and evaluating soil health could be necessary because, worldwide, many farms have suboptimal soil fertility [72]. Pest infestations significantly impact cocoa yields. Research highlights that low adoption of good farming practices combined with pest and disease attacks contribute to reduced yields. Understanding local pest populations and implementing effective management strategies are crucial for mitigating these challenges. Sustainable farming practices are essential for long-term cocoa cultivation [73]. A global review of cocoa farming systems identifies six key drivers of on-farm productivity: variety cultivated, soils, farm husbandry, farm age, abiotic factors (climate), and biotic factors (pests, diseases, weeds, parasitic plants) [74]. Integrating these factors into farming practices can improve productivity and sustainability. Additionally, agroforestry practices have positively impacted soil microbial diversity and nutrient cycling, improving soil quality. Implementing such systems can lead to more resilient and productive cocoa farms [75,76].

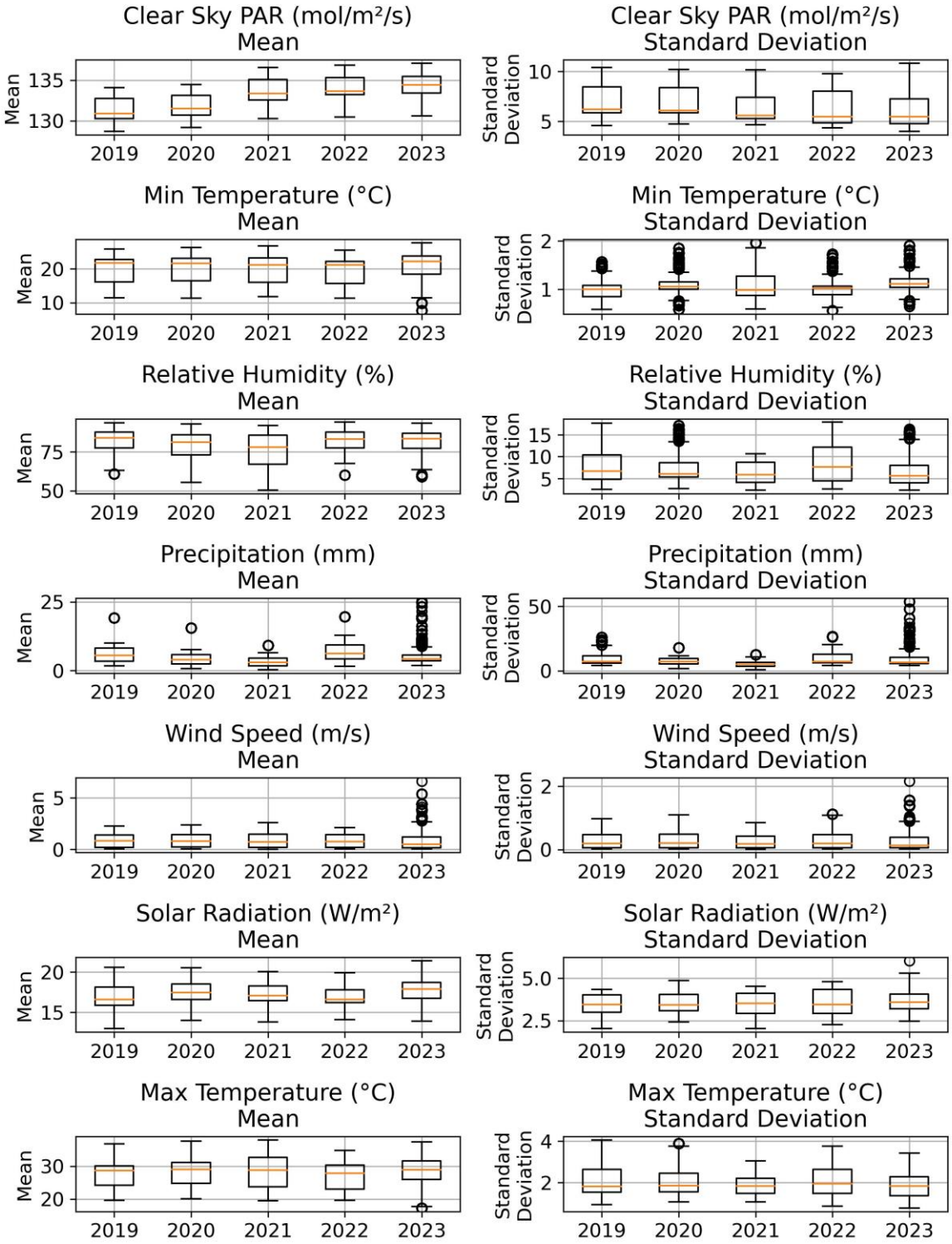

Figure 8. Boxplot for annual climatic variable distributions in cocoa cultivation regions (2019–2023).

The environmental conditions for cocoa cultivation in Colombia demonstrate a generally stable and favorable climate with nuanced regional differences that influence production potential. The temperature values, both maximum and minimum, show low variability and remain within the optimal range for cocoa growth. This stability supports critical physiological processes and reduces extreme heat or cold stress risks. The narrow spread of solar radiation and photosynthetically active radiation (PAR) indicates consistent light availability, which is crucial for photosynthesis and uniform growth cycles.

Relative humidity reflects the tropical nature of the region, with moderate variability centered around a high mean. These conditions are ideal for cocoa growth, although areas with elevated humidity may influence disease dynamics. The precipitation data reveal moderate variability, highlighting differences in rainfall distribution. Regions with lower precipitation may face water stress, while areas experiencing higher rainfall might contend with localized waterlogging.

The wind speed data suggest calm conditions with minimal variability, fostering a stable microclimate and reducing risks such as excessive evaporation or physical damage to plants. The soil moisture variables, including surface and profile wetness, exhibit moderate variability with consistent water availability in most regions. However, root zone soil wetness displays higher variability, with lower mean values indicating potential challenges in deeper soil layers, particularly in drier or semi-arid zones. Table 2 describes the average values observed in the window analysis.

Table 2. Interpretation of average environmental variable values for cocoa cultivation in Colombia.

Variable	Mean	Std	Min	25%	50%	75%	Max	Interpretation of Distribution
Max Temperature (°C)	28.18	0.72	27.12	27.76	28.58	28.71	28.74	Low variability in the annual average indicates a stable temperature regime, suitable for enzymatic processes and physiological stability. The small spatial and temporal standard deviation (mean: 1.99 °C, std: 0.10 °C) across regions suggests homogeneity in conditions. Minimal variation between areas reduces the likelihood of extreme temperature disparities.
Solar Radiation (W/m^2)	17.20	0.40	16.76	16.97	17.13	17.34	17.80	The consistent annual average reflects uniform energy availability. Spatial and temporal variability (mean std: 3.52 W/m^2, std: 0.07 W/m^2) across regions is narrow, with localized variations likely caused by transient atmospheric or geographic factors. The overall trend supports predictable growth conditions for cocoa.
Wind Speed (m/s)	0.82	0.06	0.73	0.81	0.81	0.84	0.90	Low variability in wind speed is observed in both annual averages and spatial distributions (mean std: 0.28 m/s, std: 0.02 m/s). These stable conditions prevent excessive water loss through evaporation and support the microclimatic stability necessary for consistent evapotranspiration and pollination processes.
Clear-Sky PAR ($mol/m^2/s$)	132.91	1.39	131.23	131.61	133.60	133.88	134.25	The annual average photosynthetically active radiation (PAR) values indicate highly consistent photosynthetic energy. The spatial and temporal deviations (mean std: 6.47 $mol/m^2/s$, std: 0.39 $mol/m^2/s$) suggest minimal regional variability, with slight peaks likely attributed to clearer atmospheric conditions or reduced cloud cover in specific zones.
Relative Humidity (%)	80.09	2.64	76.29	78.43	81.52	81.63	82.57	Moderate variability in the annual average is complemented by spatial and temporal deviations (mean std: 7.10%, std: 0.79%). The lower regional variability supports tropical climate uniformity, though localized humidity peaks or troughs could influence biological and fungal processes in specific cultivation areas.
Min Temperature (°C)	20.03	0.47	19.44	19.91	20.02	20.02	20.76	The narrow spread in annual averages demonstrates thermal stability crucial for metabolic recovery. The spatial and temporal variability (mean std: 1.06 °C, std: 0.06 °C) remains low, confirming minimal geographical disparities in nighttime cooling patterns, which benefits consistent development across regions.
Precipitation (mm)	5.00	1.46	3.07	4.06	5.36	5.72	6.80	Precipitation shows moderate variability, with annual averages and spatial and temporal deviations (mean std: 7.78 mm, std: 1.77 mm) indicating significant differences across regions and days. Lower values represent drought-prone areas, while higher values reflect zones of concentrated rainfall, requiring tailored water management in specific locations.
Surface Soil Wetness	0.77	0.11	0.39	0.70	0.79	0.85	0.98	Moderate annual averages with spatial and temporal deviations (mean std: 0.12, std: 0.06) indicate variability in surface moisture retention. Regions with high wetness likely benefit from greater water availability, while areas with lower values suggest transient drying or lighter soil textures.
Profile Soil Wetness	0.79	0.10	0.36	0.71	0.79	0.87	0.99	High centrality in annual averages is paired with spatial and temporal variability (mean std: 0.10, std: 0.04), reflecting dependable subsurface water availability. Outliers near the minimum may correlate with challenges in water retention across less permeable soils or drier zones.
Root Zone Soil Wetness	0.78	0.10	038	0.72	0.80	0.87	0.99	The high variability in root zone wetness (mean std: 0.10, std: 0.04) indicates significant regional disparities in deep soil moisture. Areas with lower moisture retention may face stress during dry spells, while higher values correspond to saturated zones, highlighting diverse water dynamics across the landscape.

3.1.2. Correlation Analysis

The correlation between the different environmental variables observed in Figure 9 reveals intricate relationships significantly impacting cocoa cultivation. These correlations highlight how variations in one factor can influence or predict changes in another, shaping the overall cultivation environment. Key findings include the following:

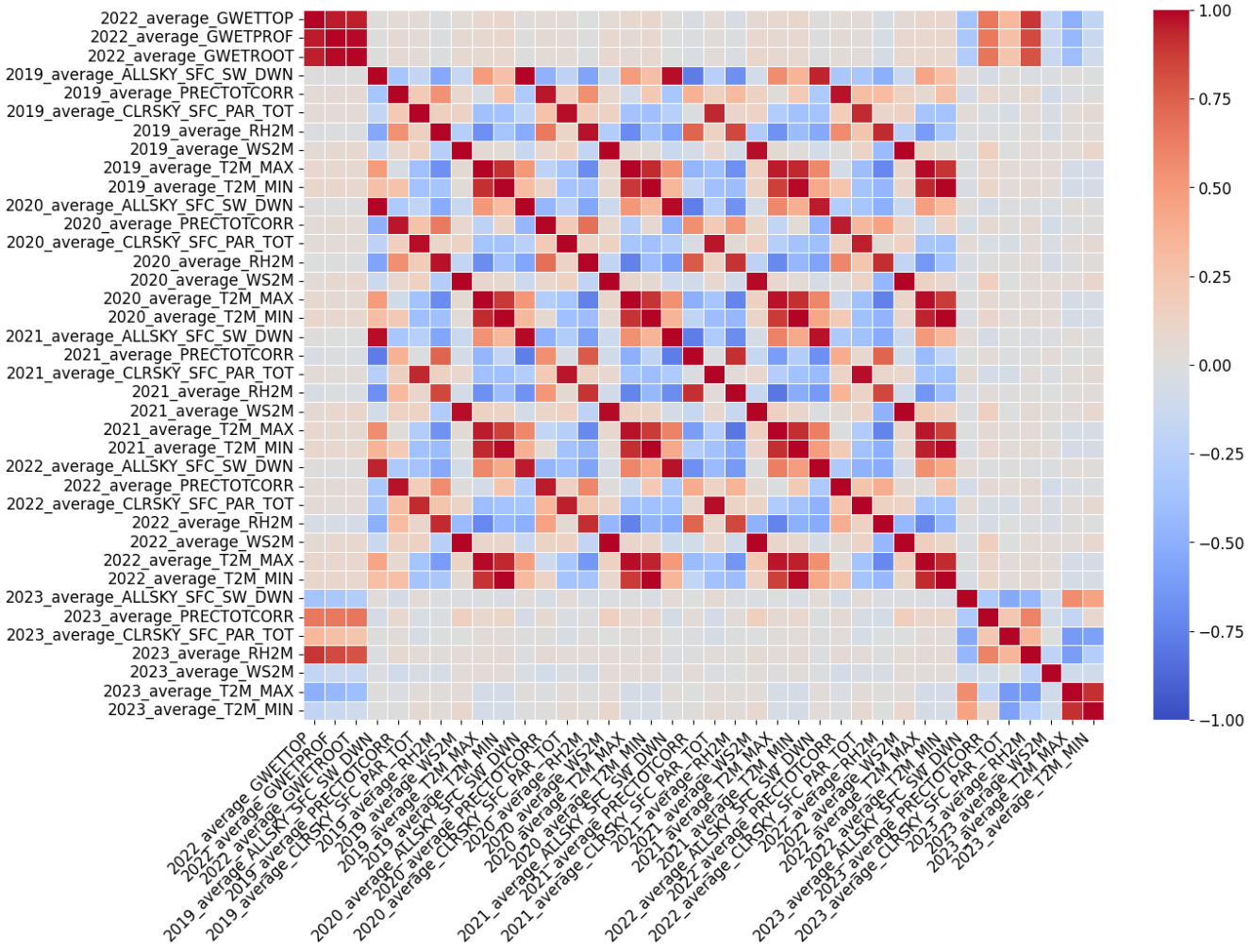

Figure 9. Correlation heatmap of meteorological averages (2019–2023).

The maximum and minimum temperatures (T2M_MAX, T2M_MIN) consistently correlate with the relative humidity (RH2M) across multiple years. This relationship suggests that as temperatures increase, humidity levels also tend to be higher, which can impact cocoa plants by increasing transpiration rates and potentially influencing pests and disease prevalence. All-sky Solar Downward Radiation (ALLSKY_SFC_SW_DWN) and precipitation (PRECTOTCORR) exhibit a significant inverse correlation, particularly noted in 2019 and 2020. These correlations suggest that higher solar radiation often coincides with lower precipitation levels. This dynamic can lead to increased water demand for cocoa plants due to higher evapotranspiration on sunny days, necessitating effective water management strategies during these periods.

Wind speed (WS2M) has a moderate to strong correlation with the maximum temperature (T2M_MAX) across the years. This correlation indicates that higher temperatures can be associated with increased wind speeds, which affect cocoa through mechanisms like enhanced evapotranspiration and potentially increased pollination or dispersion of pests. All-sky Solar Downward Radiation (ALLSKY_SFC_SW_DWN) has a high correlation with

clear-sky photosynthetically active radiation (CLRSKY_SFC_PAR_TOT), particularly in years like 2019 and 2020. This strong positive correlation underscores the critical role of sunlight in providing energy for photosynthesis, which is essential for the growth and productivity of cocoa plants.

In addition, the standard deviations of maximum and minimum temperatures in Figure 10 often show correlations, indicating that the variability in temperature from day to night can be consistent across the years. This behavior suggests a stable but potentially challenging environment for cocoa cultivation, as significant temperature swings can stress plants and affect growth cycles. The relative humidity (RH2M) shows a correlation with soil moisture variables (GWETTOP, GWETPROF, GWETROOT), particularly in recent years such as 2022 and 2023. This relationship highlights how ambient moisture levels can reflect or influence soil moisture conditions, which is crucial for maintaining adequate water availability for cocoa tree roots.

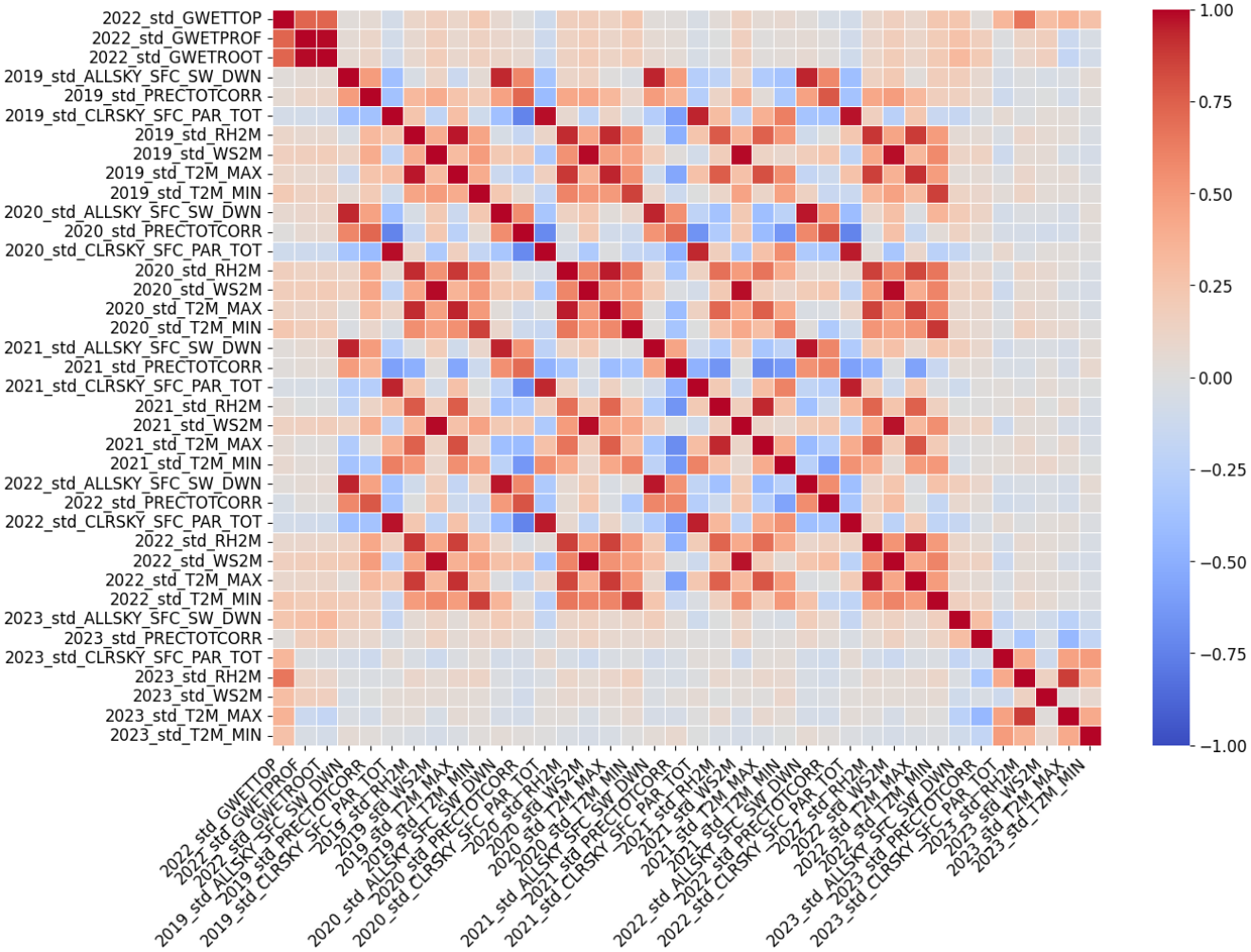

Figure 10. Correlation heatmap of meteorological variability (2019–2023).

Complementary datasets were generated to deepen the analysis of environmental variables affecting cocoa cultivation, including correlation and *p*-value matrices for the average of and variability in meteorological parameters. These datasets provide a quantitative foundation for evaluating the significance and strength of observed relationships. The standardized and average-based analyses provide a more holistic perspective, accommodating typical conditions and variability, which are critical for anticipating environmental challenges and devising resilient management strategies.

3.2. Cluster Analysis Result

The first step in this stage is to identify the optimal number of clusters or homogeneous groups. The selection of 10 clusters is grounded in a thorough examination of clustering performance, validated through the elbow method and silhouette analysis. Each method provides complementary insights that support the choice of 10 clusters as a viable structure for distinguishing environmental patterns relevant to cocoa establishment.

In Figure 11, the elbow method plots the inertia, or within-cluster sum of squares, against the number of clusters, revealing a distinct "elbow" at the 10-cluster mark. This point represents a substantial reduction in inertia (especially considering the preliminary change with fewer clusters), suggesting that increasing the number of clusters beyond this point yields diminishing returns. Specifically, after 10 clusters, the decrease in inertia becomes more gradual, indicating that additional clusters capture increasingly minor variations within the data. This inflection point is a visual signal that 10 clusters effectively balance the trade-off between data variance and clustering efficiency, capturing substantial intra-cluster homogeneity without over-fragmentation.

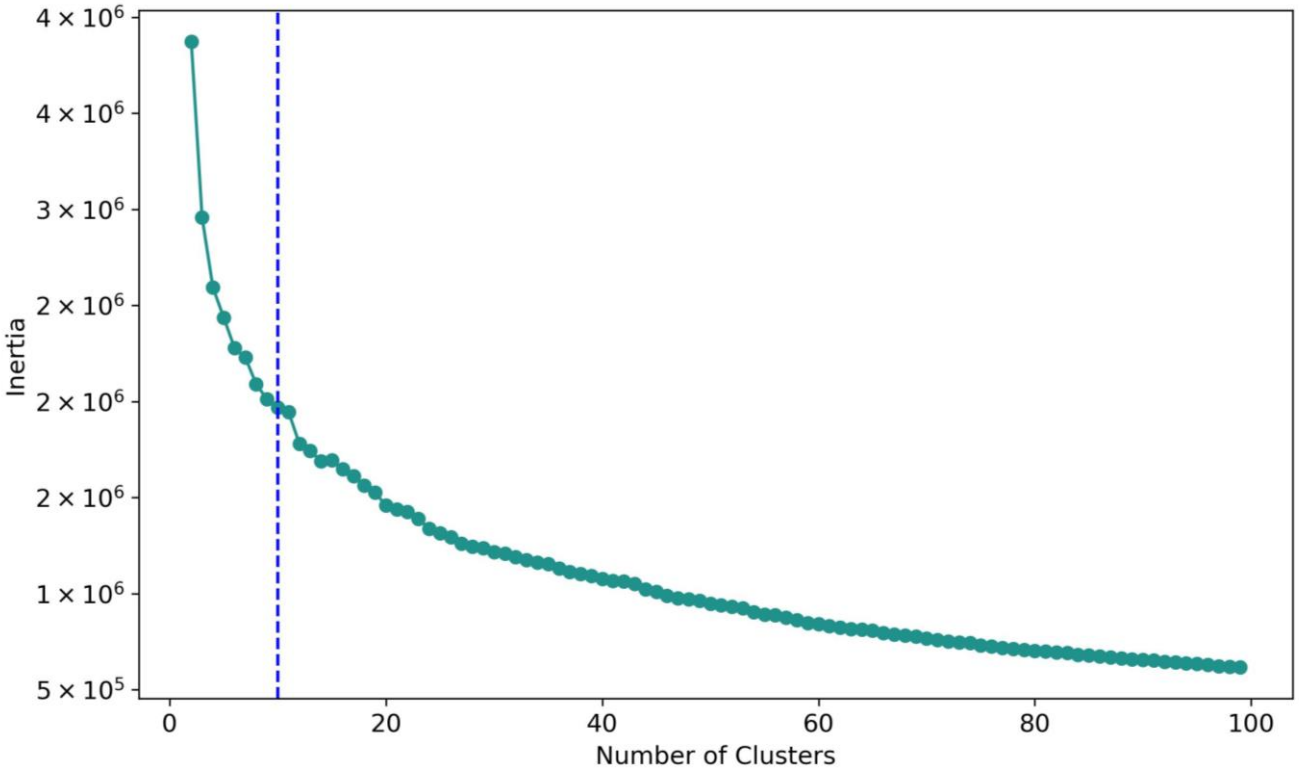

Figure 11. The elbow method indicating the slope change.

Complementing this, Figure 12 illustrates the results of silhouette analysis, which measures the cohesion and separation of clusters by assessing how well each point aligns with its assigned cluster compared to others. The average silhouette score across all clusters is approximately 0.21, as indicated by the red dashed line. Although a higher silhouette score is generally preferred, a score of 0.21 suggests that the clusters possess moderate separation, allowing practical, if not highly distinct, data partitioning. This moderate silhouette value implies that while clusters are adequately cohesive, the environmental characteristics defining each cluster may not be sharply distinct, reflecting similarities or overlapping environmental conditions across the clusters.

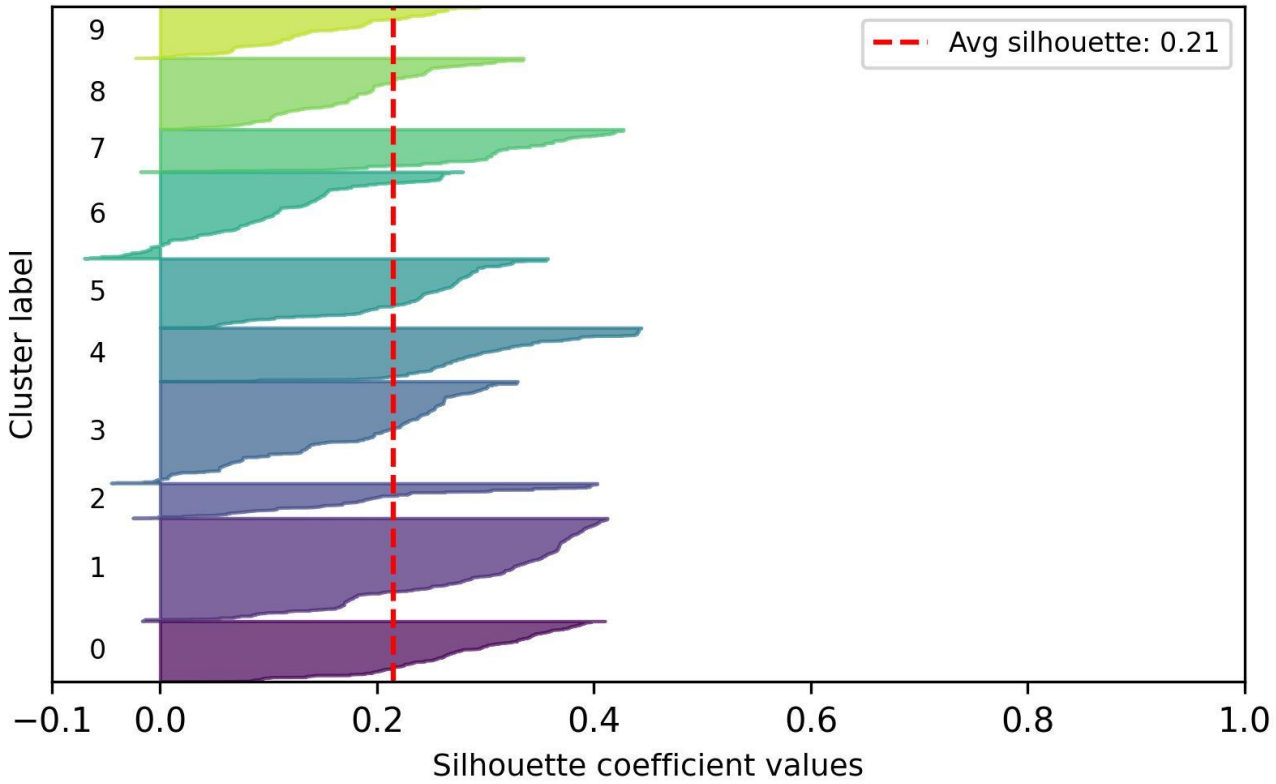

Figure 12. Density of the ten clusters.

The distribution of silhouette scores within each cluster further illustrates this point. While most clusters show predominantly positive silhouette values, the absence of strong peaks in the distribution of silhouette scores means that there are no clusters with a high concentration of members that are exceptionally cohesive and clearly separated from others. In other words, the clusters do not exhibit sharp, well-defined groupings that stand out distinctly. Instead, the distribution suggests a more moderate level of cluster definition, where groups are reasonably cohesive but not completely isolated. This balanced distribution supports the idea that the clusters, while distinct enough to capture meaningful variation, still share some environmental overlap, which may be inherent to the geographic and climatic gradients observed in Colombia.

The clusters in Figure 13 highlight distinct cocoa suitability zones across Colombia, shaped by regional climates and weather impacts. In the Andean region, moderate temperatures, high humidity, and steady rainfall provide favorable conditions for cocoa, though mountainous terrain creates microclimates with varying moisture and temperature needs. Soil conservation and moderate shading are essential to enhance growth here.

Along the Pacific coast and western Andes, intense rainfall and high humidity support cocoa but increase fungal disease risks, making agroforestry and canopy management critical for airflow and moisture control. Warmer temperatures and variable rainfall in the northern lowlands challenge water availability and heat stress. Water conservation and shading are vital to retain soil moisture and reduce plant stress during dry periods. The Amazonian and southern regions experience high humidity and consistent warmth, which benefits cocoa but requires careful temperature management to prevent heat stress. Ground cover and mulching stabilize soil temperatures, while rainwater harvesting ensures reliable moisture.

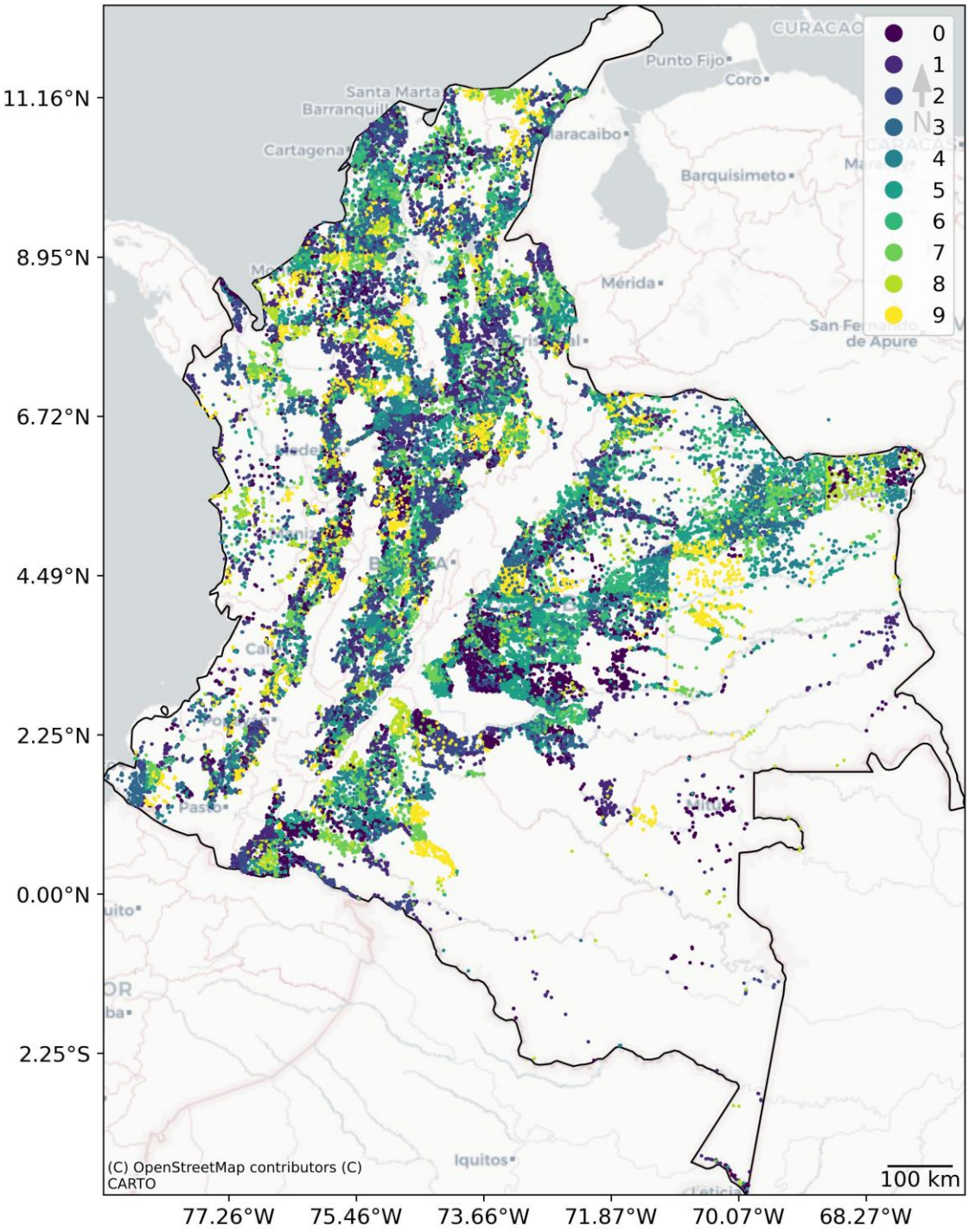

Figure 13. Regional distribution of cocoa suitability clusters in Colombia.

The analysis of environmental variables across the ten clusters in Figure 14 highlights distinct ecological patterns that influence cocoa establishment potential in Colombia. Clusters 0 and 4 stand out for their elevated soil moisture levels across all three measured depths: surface (GWETTOP), profile (GWETPROF), and root zone (GWETROOT). With averages around 0.82–0.83 in Cluster 0 and approximately 0.78 in Cluster 4, these regions maintain a high and consistent moisture profile supporting cocoa root development and nutrient uptake. In contrast, Clusters 6 and 7, showing lower soil moisture (0.74–0.75), may face drier conditions, suggesting a need for irrigation intervention to maintain soil moisture levels favorable for cocoa.

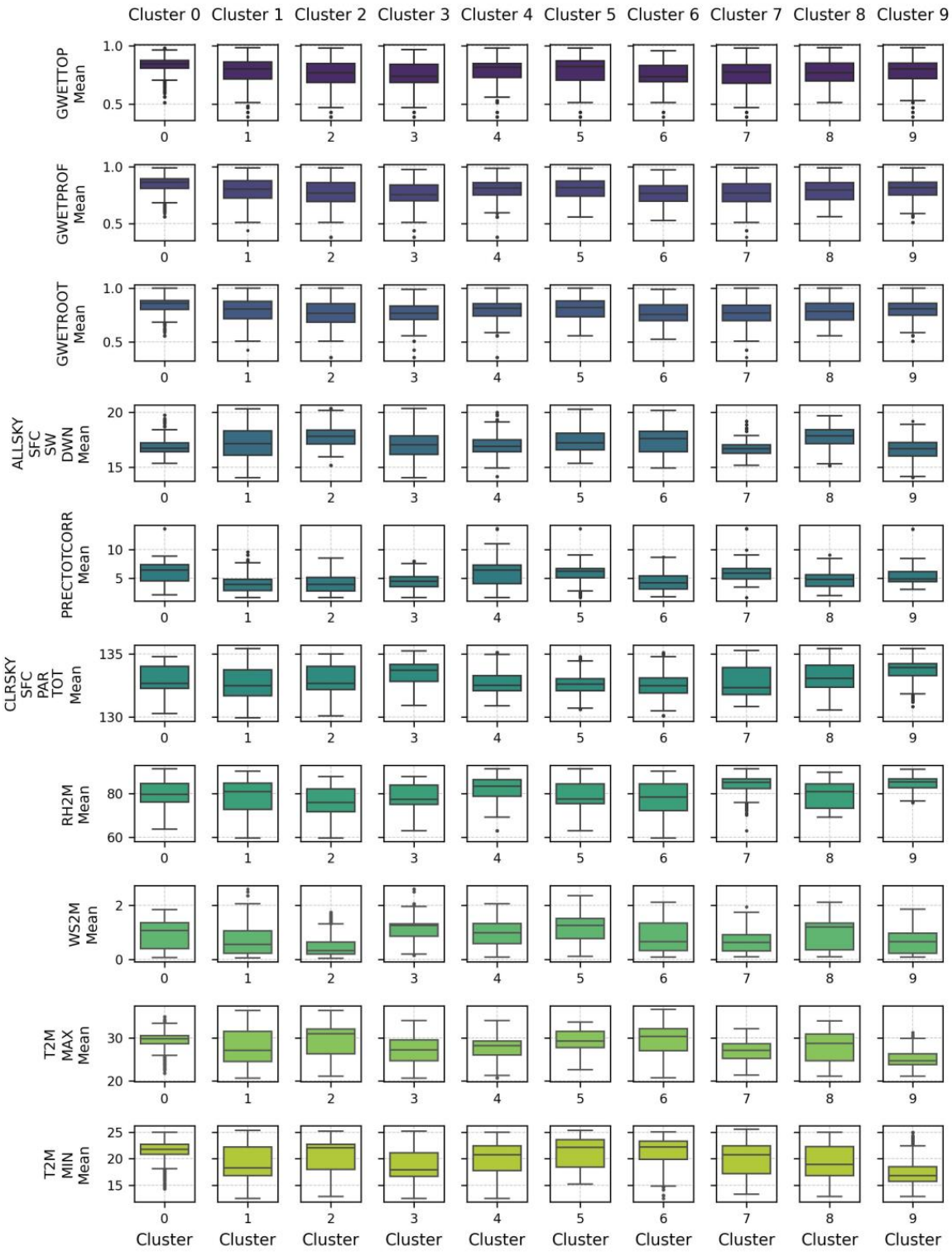

Figure 14. Boxplot comparison of environmental variables across cocoa suitability clusters in Colombia.

Solar radiation varies significantly across clusters, affecting the plant's growth rate and water requirements. Cluster 2 receives the highest solar radiation, averaging 17.86 MJ/m^2/day, which could accelerate growth due to increased photosynthetic activity but may concurrently increase evapotranspiration. In comparison, Cluster 9 has the lowest solar radiation (16.59 MJ/m^2/day), which might reduce water stress but could also slow plant metabolic processes if insufficient energy reaches the canopy. Meanwhile, photosynthetically active radiation (PAR) shows minor variability. However, Cluster 9's slightly elevated average (133.61 µmol/m^2/s) might offer an advantage in supporting

photosynthesis and growth under low light conditions, provided that soil moisture is adequately managed.

Precipitation levels further differentiate the clusters, impacting each area's water availability and suitability for cocoa cultivation. Cluster 0 receives the highest mean precipitation (6.11 mm/day), contrasting sharply with Cluster 1, where precipitation drops to 3.97 mm/day. This discrepancy suggests that Clusters 1 and 3, with lower precipitation levels, may benefit from additional water resource management, such as rainwater harvesting, to sustain crop health during dry periods.

Relative humidity patterns add another layer to the environmental profile. Clusters 7 and 9 exhibit high humidity levels (84.1% and 84.6%, respectively), which could reduce drought risk by lowering plant transpiration rates. However, elevated humidity might also predispose these regions to fungal diseases, a common concern in cocoa production under humid conditions. On the other hand, Cluster 2, with a lower average humidity of 76.6%, might face drier air conditions, which could lessen the prevalence of humidity-related diseases but may require increased irrigation to mitigate soil moisture loss.

Wind speed provides insights into surface drying and potential evaporation rates, with Clusters 3 and 5 experiencing higher average wind speeds (1.09–1.13 m/s). These conditions could lead to accelerated soil drying, impacting the water retention necessary for sustained cocoa growth. In contrast, Cluster 2's notably lower wind speed (0.49 m/s) may aid in preserving soil moisture, supporting cocoa growth with less evaporation-related moisture loss.

Temperature variations between clusters add further complexity to cocoa suitability. The maximum temperature in Cluster 6 (29.7 °C) suggests a warmer environment, conducive to faster metabolic rates, but it might also heighten water requirements. Conversely, Cluster 9's lower average maximum temperature (25.2 °C) indicates a cooler environment, which could slow metabolic processes and growth rates. Minimum temperatures also vary, with Cluster 9 experiencing the lowest (17.7 °C), potentially affecting nighttime respiration. In comparison, Cluster 6's higher minimum temperature (21.2 °C) maintains warmer nighttime conditions that could accelerate plant processes but may also intensify moisture loss.

Clusters 0, 4, and 6 appear favorable for cocoa establishment due to their soil moisture, radiation, and temperature balance. However, clusters like Cluster 2, with lower humidity and higher solar radiation, would require careful water management to prevent plant stress. Cluster 9, with its cooler, humid environment, might suit cocoa but demands monitoring for humidity-related diseases.

3.3. Model Performance Comparison

The algorithms performed well across both datasets—one unbalanced (different number of points in each category) and one balanced (same quantity of points corresponding with each category), using the SMOTE or Synthetic Minority Over-sampling Technique to balance the classes—in forecasting cocoa cultivation aptitude in Colombia, with random forest emerging as the most effective for classifying low-, medium-, and high-aptitude regions. Solving the classification problem with unbalanced categories, random forest achieved the highest accuracy (94.11%) on the balanced dataset, followed closely by decision tree (92.97%) and SVM (85.93%). Logistic regression showed the weakest performance, with an accuracy of 54.44%. Table 3 summarizes the algorithms' performances. Figures 15 and 16 present confusion matrices for both the balanced and unbalanced datasets.

Table 3. Classification report.

Model	Dataset	Accuracy	Precision (Avg)	Recall (Avg)	F1-Score (Avg)
Decision Tree (Depth 20)	Unbalanced	93.62%	93.70%	93.62%	93.62%
	Balanced	92.97%	93.03%	92.97%	93.68%
Random Forest (Depth 15)	Unbalanced	93.24%	93.40%	93.24%	93.23%
	Balanced	94.11%	94.14%	94.11%	94.11%
Neural Network	Unbalanced	91.37%	91.45%	91.37%	91.34%
	Balanced	91.14%	91.23%	91.14%	91.24%
Logistic Regression	Unbalanced	54.44%	54.13%	54.44%	53.71%
	Balanced	54.37%	54.27%	54.37%	54.57%
SVM (RBF Kernel)	Unbalanced	85.93%	85.99%	85.93%	85.86%
	Balanced	86.99%	86.99%	86.99%	86.99%

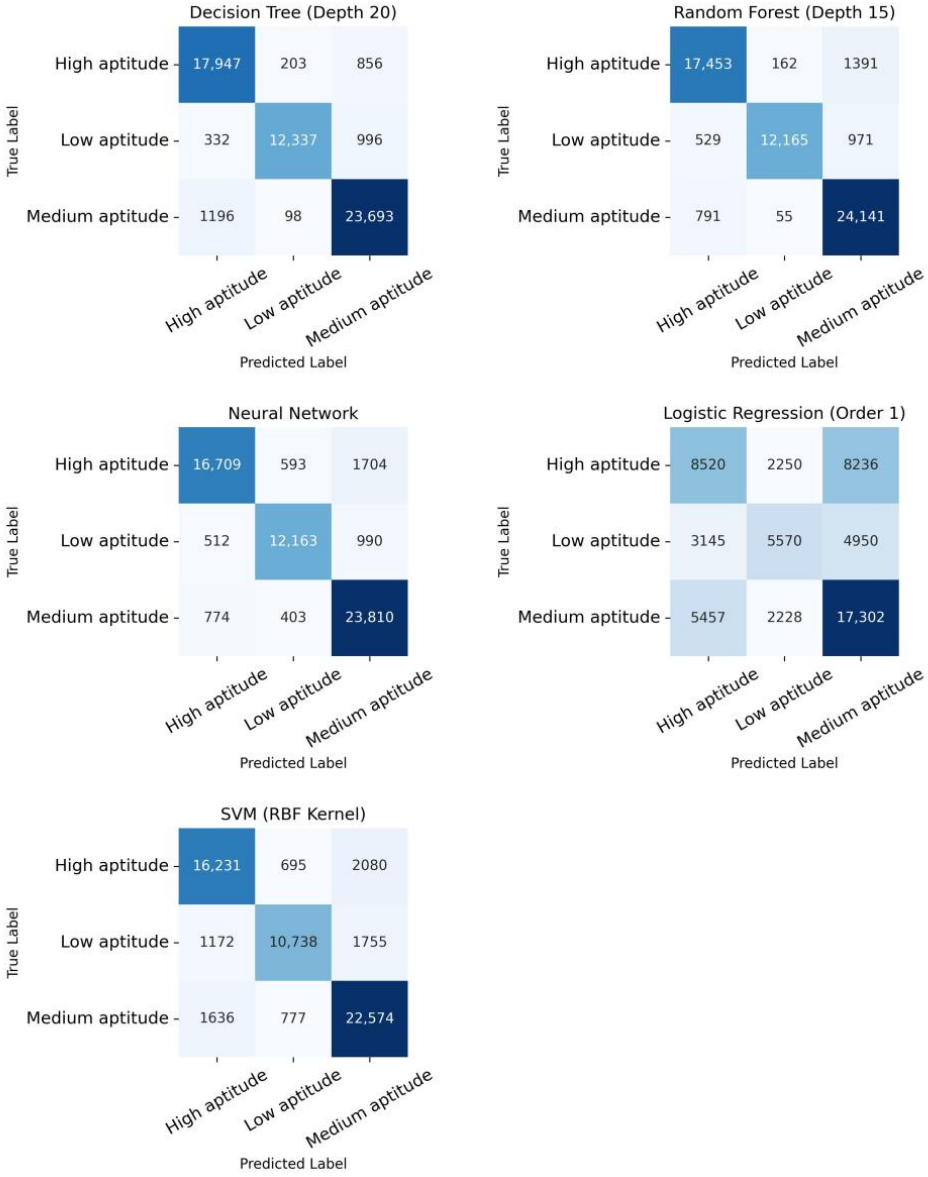

Figure 15. Confusion matrices: unbalanced categories.

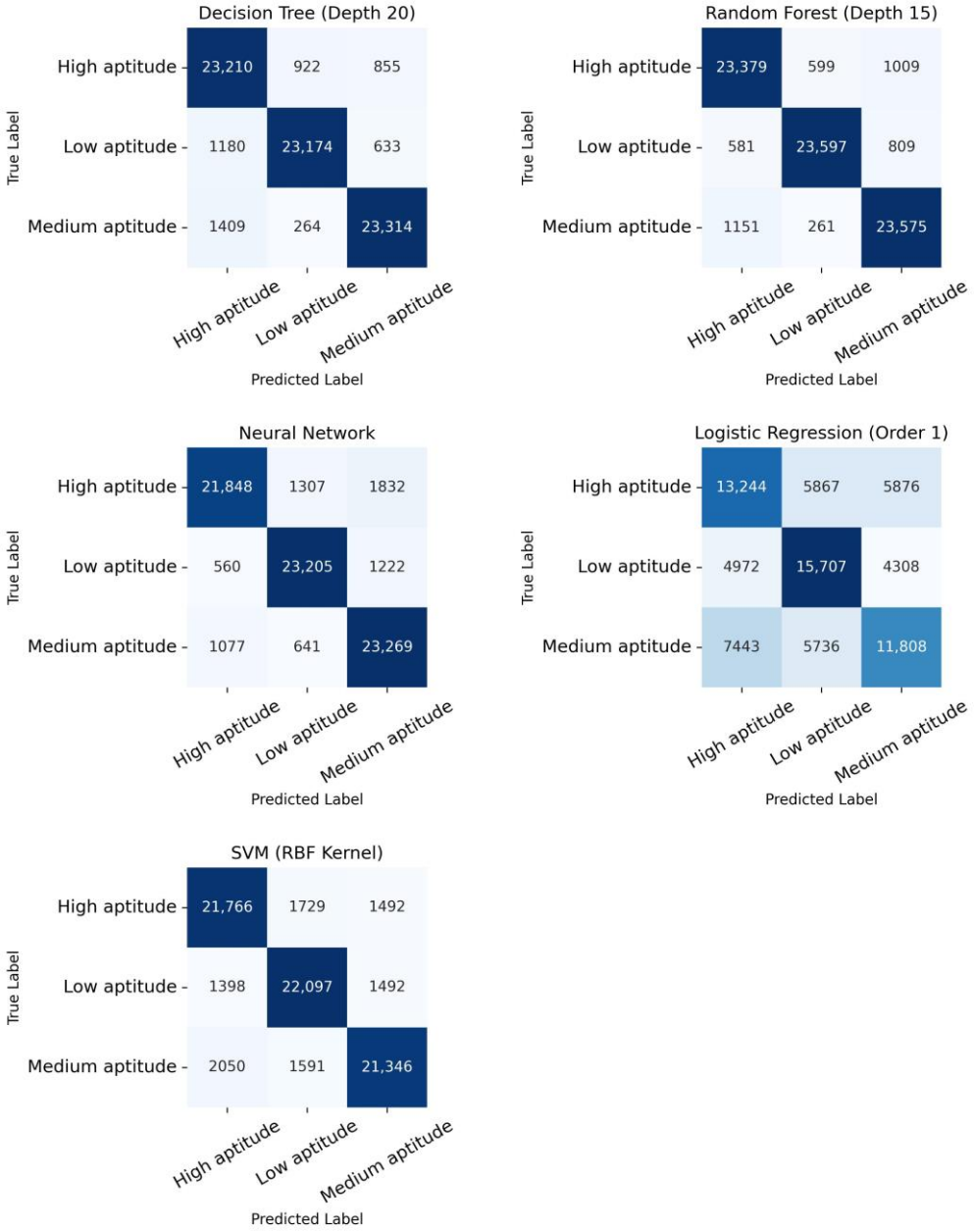

Figure 16. Confusion matrices: balanced categories using SMOTE.

3.3.1. Performance with Unbalanced Data

Decision Tree: This model demonstrates strong classification performance, with a high overall accuracy of 93.62%. The precision rates are substantial across all aptitude levels: 92.15% for high, 97.61% for low, and 92.75% for medium aptitude. The recall rates are similarly balanced, particularly excelling in the high (94.43%) and medium (94.82%) categories. This consistency results in a high weighted F1-score of 93.62%, indicating the decision tree model's robustness in scenarios where precision and recall are equally important. This model is particularly reliable for balanced class identification with minimal bias, making it suitable for diverse aptitude classifications.

Random Forest: With an impressive accuracy of 93.24% and the highest precision for low aptitude at 98.25%, random forest proves exceptional for maintaining high recall and precision, notably with a low tendency for misclassification. Its F1-score is strong across all categories, especially for low aptitude (93.41%), highlighting its effectiveness and

inconsistent classification. This model is particularly effective in accurately differentiating classes with high precision, providing a reliable solution for nuanced classification tasks.

Neural Network: Achieving an overall accuracy of 91.37%, the neural network model displays competitive performance but slightly lower precision for high aptitude (92.85%) than the decision tree and random forest models. While it has strong recall and F1-scores—particularly for medium aptitude (92.48%)—its slightly lower overall accuracy suggests it may be less effective in high-precision and differentiation scenarios. This model remains a valuable choice for tasks requiring a blend of high recall and general adaptability across categories. However, it does not match the accuracy of the decision tree or random forest models.

Logistic Regression: This model exhibits substantial limitations, with a significantly lower overall accuracy of 54.44%. It has lower precision and recall scores across all aptitude categories, underscoring a high risk of misclassification. The model's limited performance in distinguishing aptitude levels suggests it is less suitable for scenarios where accuracy and reliable differentiation are essential. This highlights the need for either further adjustments or an alternative model when accuracy is a priority.

SVM: This model performs well, achieving an accuracy of 85.93% and strong precision across all categories. While the recall for low aptitude is slightly lower at 78.58%, SVM maintains a solid F1-score, especially for medium aptitude (87.84%), ensuring reliable performance in balanced classifications. Although slightly prone to underestimation in the low-aptitude category, the model is well-suited for scenarios requiring balanced classification with reasonable precision and recall across classes.

The performance differences among the models can be attributed to their inherent structures and capacities to manage data complexity. Random forest emerged as the most accurate model, thanks to its ensemble nature, which averages errors across multiple decision trees, thus capturing complex patterns while minimizing overfitting due to diverse trees, which helps smooth out noise and outliers [77]. In contrast, the decision tree model, with a maximum depth of 20, performed well but slightly underperformed RF due to the limitations of featuring only a single tree, which is more susceptible to overfitting.

The ANN, designed with a single hidden layer of 100 neurons, demonstrated strong performance, leveraging its capacity to model non-linear relationships. However, its accuracy was slightly lower than RF's, suggesting that additional layers or neurons would strengthen its capabilities. Nevertheless, single-hidden-layer ANNs do not experience overfitting when overtraining is avoided by cross-validation [78]. SVM with an RBF kernel offered reasonable accuracy. However, its reliance on kernel choice and parameter tuning may limit its ability to fully capture the data's complexity and make it sensitive to overfitting [79].

Finally, logistic regression exhibited the lowest performance, primarily due to its linear structure, which needs to be improved to capture the intricate, non-linear patterns in the environmental variables. Overall, the superior performance of random forest highlights the advantages of ensemble approaches. At the same time, the ANN and SVM provide robust alternatives for non-linear modeling tasks, albeit with certain limitations in their configurations.

3.3.2. Performance with Balanced Data

The performance of the models in the balanced dataset (balance) mirrors their performance in the unbalanced case, indicating that these models maintain their robustness irrespective of dataset balance: decision trees and random forests continue to show high accuracy and precision, which is critical for considering areas suitable for cocoa cultivation. Their consistency across both datasets highlights their suitability for varied data scenarios.

The neural networks, logistic regression, and SVMs also display similar metrics in the balanced dataset, with the neural networks and SVMs providing reasonable alternatives for applications requiring high recall rates and robust classification. Considering the need to avoid underestimating cocoa cultivation aptitude, the random forest model is the most reliable, given its high precision, recall, and F1-scores. It is especially accurate in identifying low-aptitude areas. It demonstrates the capacity to minimize critical misclassifications across balanced and unbalanced datasets.

3.4. Key Predictors of Cocoa Aptitude

This analysis revealed that the random forest algorithm emerged as the most effective model for classifying areas into low-, medium-, and high-cocoa cultivation aptitudes. The detailed examination of key variables cumulatively explaining 80% of the variance in cocoa suitability highlights the intricate interplay of climatic and soil factors necessary for optimal cocoa production (see Figure 17). Starting with the minimum temperature (2023_std_T2M_MIN) as the most influential variable, it emphasizes the thermal sensitivity of cocoa, where slight deviations can significantly affect crop viability. This factor, combined with clear-sky photosynthetically active radiation (2023_std_CLRSKY_SFC_PAR_TOT), illustrates how essential sunlight exposure and suitable temperature ranges are for maximizing photosynthesis and, ultimately, cocoa bean quality.

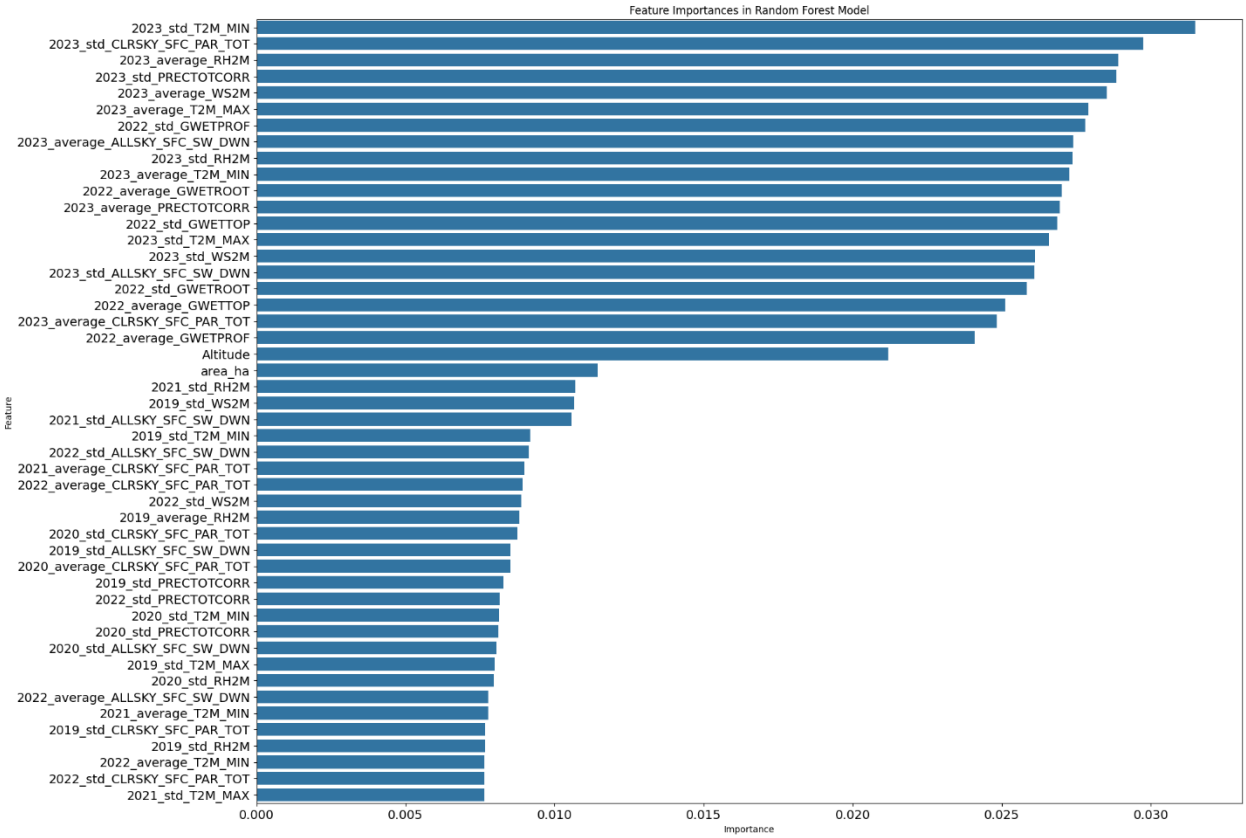

Figure 17. Parameters that account for 80% of the variance.

The importance of relative humidity (2023_average_RH2M) and precipitation variability (2023_std_PRECTOTCORR) indicates that water availability, both in terms of atmospheric moisture and soil moisture consistency (2022_std_GWETPROF, 2022_average_GWETROOT), plays a crucial role in determining areas suitable for cocoa cultivation. These variables suggest that maintaining a balance in moisture levels is key to improving

growth conditions and is a powerful tool in preventing diseases and providing reassurance about the health of cocoa crops.

Moreover, wind speed (2023_average_WS2M, 2022_std_WS2M) is a critical factor affecting cocoa farms' evapotranspiration rates and pollination. This phenomenon accentuates the need for strategic farm placement to harness or shield from wind effects based on local climatic conditions. Indeed, constant and intermittent wind exposure at speeds of 2.5, 3.5, and 4.5 m/s for 3, 6, and 12 h significantly reduced photosynthetic rates, stomatal conductance, transpiration, and water use efficiency in mature cocoa leaves, with young leaves exhibiting greater sensitivity to mechanical stress and resulting damage [80].

Additionally, solar radiation (2023_average_ALLSKY_SFC_SW_DWN, 2022_average_ALLSKY_SFC_SW_DWN) directly correlates with the potential for energy capture by cocoa plants, influencing growth rates and flowering times. By cross-referring environmental variables, researchers could picture scenarios for enhancing cocoa aptitude and crop establishment strategies, corroborating that optimal cultivation conditions require a fine balance of temperature, moisture, solar radiation, and wind. In the case of Colombia, precipitation is not a usual variable to control (e.g., using irrigation approaches) due to the relatively high precipitation in many regions of the country.

3.5. Per-Cluster Model Performance

Examining performance outcomes is essential when considering whether to apply a single model to the entire dataset or divide the data into clusters and apply distinct models to each one of them (see Table 4). By comparing model accuracies in both the whole dataset and the clustered subsets, it can be better understood how clustering affects predictive accuracy and the overall utility of machine learning models.

Table 4. Comparative matrix of per-cluster model performance.

Cluster	Random Forest (Depth 15)	Neural Network	Decision Tree (Depth 20)	SVM (RBF Kernel)	Logistic Regression
0	94.69%	92.54%	92.51%	88.32%	76.31%
1	92.79%	91.01%	90.37%	86.27%	66.63%
2	92.15%	89.49%	90.13%	85.34%	65.71%
3	92.98%	90.77%	91.30%	82.74%	67.08%
4	96.91%	96.49%	95.09%	93.21%	75.62%
5	94.28%	93.31%	92.08%	88.04%	76.74%
6	94.35%	93.15%	93.08%	88.20%	72.85%
7	91.79%	89.82%	90.35%	85.94%	71.77%
8	94.61%	93.62%	92.52%	87.53%	78.01%
9	90.92%	88.76%	88.96%	83.14%	72.93%

The performance of various models on the entire dataset shows that random forest achieves the highest accuracy at 93.62%. Neural networks and decision trees also perform reasonably well, with accuracies of 91.37% and 93.62%, respectively. SVM follows with an accuracy of 85.93%, while logistic regression significantly lags with only 54.44%. This disparity suggests that applying a single model across a heterogeneous dataset may not be optimal for capturing the diverse patterns present.

When examining the results of clustered data, it becomes clear that dividing the dataset into clusters provides a substantial performance boost, particularly for simpler models like logistic regression. Logistic regression achieves a much higher accuracy in several clusters than when applied to the whole dataset. For instance, in Cluster 0, its accuracy rises to 76.31%, while it improves to 78.01% in Cluster 8 and 76.74% in Cluster 5, respectively.

These improvements indicate that clustering isolates homogeneous zones, where simpler models perform much better, validating the homogeneity in the clusters.

For complex models such as random forest and neural networks, clustering also leads to improvements, although the differences are less pronounced than with logistic regression. Random forest's accuracy, which is already high when applied to the entire dataset, consistently increases when applied to clusters. In Cluster 40, for example, random forest achieves an accuracy of 96.91%, a notable improvement over its whole-data accuracy of 93.62%. This pattern is seen in other clusters, such as Cluster 0 (94.69%) and Cluster 8 (94.61%), highlighting that clustering enables more granular models to excel by focusing on localized data patterns.

The improvements in the performance of the simpler logistic regression model when applied to clustered data are a testament to the potential of this approach. Logistic regression, which struggles when applied to the entire dataset, significantly benefits when trained on clusters that capture consistent weather, soil, or environmental conditions. The homogeneity within clusters allows logistic regression to represent linear relationships better, leading to better performance in zones with similar conditions. This success story of logistic regression on clustered data instills confidence in its potential.

Additionally, clustering helps avoid overfitting, as models applied to more homogeneous subsets of the data generalize better to specific regions within the dataset. A single model applied to the entire data may need help with generalization, leading to overfitting in certain areas while underperforming elsewhere. By applying distinct models to clusters, the overall predictive performance improves as each model is better suited to the specific characteristics of its respective subset.

Despite the added complexity of managing multiple models across clusters, the performance gains justify this approach. Models like logistic regression, which perform poorly on the whole dataset, become viable options when applied to clustered data, achieving accuracies above 0.70 in several cases. Moreover, complex models like random forest and neural networks, while already strong on the whole dataset, further improve when applied to specific clusters. The clustering process successfully segments the data into more interpretable and manageable portions, allowing each model to focus on more specialized zones.

Ultimately, the analysis demonstrates that dividing the data into clusters and applying different models for each subset yields superior results compared to applying a single model to the entire dataset. Clustering allows the models to specialize and perform better by capturing homogeneity in the data. This approach enhances performance across all models, especially for logistic regression, and offers significant improvements for random forest and neural networks.

4. Discussion

4.1. Summary of Main Findings

This research integrates diverse datasets, including NASA POWER environmental and topographical data, elevation data from APIs, and Colombian agricultural sources, to address critical gaps in the existing studies [1–3,32,33,37]. The findings highlight the substantial impact of environmental factors on cocoa cultivation, offering data-driven insights to enhance productivity and sustainability amidst changing climate conditions [12,24,34]. The clustering approach proved especially effective in isolating regional patterns, reducing intra-cluster variability, and increasing prediction accuracy. This methodological refinement significantly improved the performance of simpler models, such as logistic regression, which showed better accuracy within clusters compared to the entire dataset. Such integrated approaches underscore the importance of effectively selecting the number of clusters.

The choice of the number of clusters played a critical role in these results. The elbow method (Figure 11) was used to determine the optimal number of clusters, providing a balance between reducing intra-cluster variability and maintaining computational efficiency. While this approach yielded logical results, it is worth noting that different cluster configurations may lead to variations in homogeneity and subsequent predictions. This introduces some uncertainty in the findings, especially in regions with highly dynamic environmental conditions. Recognizing this complexity helps frame key methodological contributions.

This study's methodological contributions are particularly noteworthy in using clustering to improve model performance without increasing complexity. The clustering approach effectively isolated regional patterns while reducing intra-cluster variability and improved model accuracy, particularly for simpler algorithms like logistic regression. By creating clusters that grouped similar environmental conditions, this study tailored models to these localized conditions, enabling a more refined analysis and facilitating specific agricultural recommendations. Assessing cluster quality further illuminates these refinements.

The silhouette coefficient was calculated to measure the cohesion and separation of data points within clusters to evaluate clustering quality. The average silhouette score of 0.21 (Figure 12) indicates moderate clustering quality. This suggests that the clusters captured regional environmental patterns reasonably well; however, the observed overlap in silhouette values across some clusters introduced uncertainty in the reliability of region-specific predictions. This overlap may reflect inherent similarities in environmental conditions between certain regions, resulting in less distinct boundaries. Incorporating more granular environmental data or refining the clustering approach could further enhance the precision of agricultural recommendations. Model-specific analyses bolster these insights by identifying pivotal environmental variables.

Random forest analysis was pivotal in identifying the key variables influencing cocoa suitability. This approach achieved a high overall accuracy of 94.11% and balanced the precision, recall, and F1-scores across different suitability levels, thereby validating the model's reliability. Random forest's ability to handle non-linear relationships and its feature importance analysis provided valuable insights, particularly in highlighting which environmental factors—such as minimum temperature and soil moisture variability—were most influential in determining cocoa suitability. Integrating these findings within an ensemble approach amplifies predictive robustness.

The ensemble model was crucial for addressing the complexity of cocoa suitability predictions by integrating multiple machine learning algorithms. Table 1 outlines the key configurations used for each algorithm within the ensemble. Rather than relying on extensive parameter tuning, the ensemble's strength lies in combining the predictive power of models such as decision trees, random forests, and neural networks, thus improving accuracy and robustness. This strategy mitigated limitations inherent in individual algorithms, such as overfitting in complex models or reduced interpretability in simpler ones. By leveraging the complementary strengths of these algorithms, the ensemble model provided reliable predictions across diverse environmental conditions without requiring extensive parameter-specific adjustments. Notably, simpler algorithms also benefit when tailored to cluster-based contexts.

Furthermore, this study revealed notable differences in model performance based on a model's complexity and data context. Simpler models like logistic regression demonstrated limited accuracy when applied to the full dataset, achieving only 54.44%. However, their accuracy improved significantly within specific clusters, often exceeding 70%. This finding emphasizes the value of clustering in improving the interpretability and applicability of simpler models by focusing them on more homogeneous data subsets. This approach increases prediction accuracy and provides more actionable insights applicable in practical

agricultural settings. Such performance gains reflect the synergy between clustering and detailed environmental insights.

Further analysis revealed that environmental factors explained 80% of the variance in suitability predictions, validating the model's reliability and scientific utility, as demonstrated in the model performance evaluation. Additionally, confusion matrix analysis demonstrated high recall for the "low" suitability category, which is crucial for resource optimization by minimizing false negatives. However, occasional misclassifications between "medium" and "high" suitability zones indicate that more granular environmental data could further refine predictions and enhance reliability. These evaluations strengthen the model's validity while highlighting areas for improvement.

The practical implications of this study are particularly relevant to cocoa farmers, policymakers, and agricultural planners. Identifying good agricultural practices based on localized climatic conditions can help mitigate risks associated with water stress and disease prevalence, thereby increasing resilience to climate change and promoting sustainable cocoa production. This study suggests several targeted practices based on environmental conditions:

- Shading and Sunlight Management: In regions with high solar radiation (ALLSKY_SFC_SW_DWN), agroforestry systems that integrate tall, leafy trees can provide shade, reducing sunlight intensity and preventing heat stress in cocoa plants. In areas with high clear-sky radiation (CLRSKY_SFC_PAR_TOT), using shade cloth during the early stages of cocoa growth, especially during peak sunlight, can help enhance photosynthesis without causing stress.
- Water Management Techniques: In regions with low precipitation (PRECTOTCORR), rainwater harvesting systems can ensure water availability during dry spells, maintaining necessary soil moisture levels. For areas with fluctuating surface and root zone wetness (GWETTOP and GWETROOT), contour planting and mulching can enhance water infiltration and retention, stabilizing moisture levels for cocoa plants.
- Soil Fertility and Moisture Conservation: In environments with low soil fertility (GWETPROF), organic mulching and cover cropping can improve soil structure, increase nutrient content, and retain moisture, ensuring optimal root zone wetness. Organic compost application and soil turning can also enhance the water-holding capacity of soils with high variability in root zone moisture.
- Pest and Disease Management: In high-humidity regions (RH2M), sanitation and pruning techniques can reduce moisture in the cocoa canopy, minimizing fungal infection risks. Biological control methods are recommended to manage pest populations without exacerbating humidity-related issues. Establishing windbreaks can protect cocoa plants from wind damage in regions with variable wind speeds (WS2M).
- Temperature and Wind Control: In areas with high maximum and low minimum temperatures (T2M_MAX and T2M_MIN), thermal insulation techniques such as planting ground cover and using organic mulching can moderate soil temperatures, protecting roots from extreme fluctuations that could affect growth or increase disease susceptibility.
- Enhancing Pollination: In regions with high clear-sky radiation (CLRSKY_SFC_PAR_TOT) but low natural pollinator presence, promoting the presence of *Forcipomyia* (flower flies) can enhance pollination efficiency, ensuring cocoa plants benefit optimally from sunlight through effective pollination.

Table 5 aligns each cluster's specific environmental conditions with tailored agricultural practices recommendations.

Table 5. Recommended environmental management practices for cocoa establishment across clusters in Colombia.

Cluster	Key Environmental Conditions	Recommended Practices	Specific Application
0	- High soil moisture (GWETTOP, GWETPROF, GWETROOT). - High precipitation (PRECTOTCORR). - High maximum and minimum temperatures (T2M_MAX, T2M_MIN).	- Pest and Disease Management. - Temperature Control.	- Implement sanitation and pruning techniques to reduce moisture in the cocoa canopy, minimizing fungal infection risks. - Use organic mulching and ground covers to moderate soil temperatures, protecting roots from extreme fluctuations.
1	- Low precipitation (PRECTOTCORR). - Low minimum temperatures (T2M_MIN).	- Water Management Techniques. - Soil Moisture Conservation. - Temperature Control.	- Install rainwater harvesting systems to ensure water availability during dry spells. - Apply mulching and cover cropping to retain soil moisture. - Use organic mulching to insulate the soil against low temperatures thermally.
2	- High solar radiation (ALLSKY_SFC_SW_DWN). - High maximum temperatures (T2M_MAX).	- Shading and Sunlight Management. - Temperature Control. - Enhancing Pollination.	- Integrate tall, leafy trees to provide shade, reducing sunlight intensity and preventing heat stress. - Use mulching to moderate soil temperatures. - Promote the presence of *Forcipomyia* (flower flies) to enhance pollination efficiency.
3	- Variable wind speeds (WS2M). - Average precipitation and temperatures.	- Pest and Disease Management. - Soil Fertility and Moisture Conservation.	- Establish windbreaks to protect cocoa plants from wind damage. - Implement organic mulching to improve soil structure and retain moisture.
4	- High relative humidity (RH2M). - High precipitation (PRECTOTCORR).	- Pest and Disease Management. - Water Management.	- Conduct sanitation and pruning to reduce moisture in the cocoa canopy, minimizing fungal infection risks. - Ensure adequate drainage to prevent waterlogging.
5	- High maximum and minimum temperatures (T2M_MAX, T2M_MIN). - High solar radiation (ALLSKY_SFC_SW_DWN).	- Shading and Sunlight Management. - Temperature Control.	- Use shade cloths during early growth stages to improve photosynthesis without causing stress. - Apply organic mulching and ground covers to moderate soil temperatures.

Table 5. *Cont.*

Cluster	Key Environmental Conditions	Recommended Practices	Specific Application
6	- High solar radiation (ALLSKY_SFC_SW_DWN). - High maximum and minimum temperatures (T2M_MAX, T2M_MIN).	- Shading and Sunlight Management. - Water Management Techniques.	- Integrate tall, leafy trees to provide shade and reduce sunlight intensity. - Apply mulching to enhance water retention in the soil.
7	- High relative humidity (RH2M). - Low minimum temperatures (T2M_MIN). - Low wind speeds (WS2M).	- Pest and Disease Management. - Temperature Control.	- Perform sanitation and pruning to reduce canopy moisture in cocoa plants. - Use mulching to protect roots from low temperatures by maintaining soil warmth.
8	- High solar radiation (ALLSKY_SFC_SW_DWN). - Variable wind speeds (WS2M).	- Shading and Sunlight Management. - Wind Control.	- Provide shade to reduce light intensity and prevent heat stress. - Establish windbreaks to protect plants from wind damage.
9	- Low maximum and minimum temperatures (T2M_MAX, T2M_MIN). - High relative humidity (RH2M).	- Temperature and Wind Control. - Pest and Disease Management.	- Implement thermal insulation techniques such as mulching and ground covers to moderate soil temperatures. - Conduct sanitation and pruning to reduce canopy moisture, minimizing fungal infection risks.

4.2. Comparison with Previous Studies

The results of this study align with the existing literature emphasizing the significant impact of climatic factors, particularly humidity and temperature, on cocoa yield. Cocoa is especially highly sensitive to its surrounding climatic conditions, with temperature, humidity, and wind speed emerging as critical determinants of its growth and productivity. The optimal temperature range for cocoa lies between 18 and 32 °C, where it maintains efficient physiological processes [25]. Temperatures below 15 °C can reduce yields and hinder development, while excessive heat exacerbates evapotranspiration, induces water stress, and affects overall plant vigor. Similarly, humidity is pivotal, as cocoa requires consistently high moisture levels to thrive. Low humidity levels can heighten the vapor pressure deficit, reducing photosynthetic efficiency and imposing significant physiological stress [50]. These alignments underscore the critical interplay of climate variables in cocoa yield.

Furthermore, wind speed is a less apparent yet vital factor. Strong winds can physically damage cocoa trees, particularly young and fragile plants, while increasing evapotranspiration, compounding water loss and stress [26]. These climatic elements interact synergistically, underscoring the necessity of precise environmental management in cocoa cultivation to optimize yield and ensure sustainable production practices. Previous studies have also reported that temperature extremes, in conjunction with high humidity, contribute to disease prevalence in cocoa cultivation [24]. Such multifactorial interactions highlight the need for region-specific approaches.

This research contributes uniquely by incorporating an ensemble clustering-based model that tailors cocoa suitability classification to specific regional conditions, offering a more localized and precise analysis. This approach sets this study apart from previous work by increasing prediction accuracy and providing actionable insights for smallholder

farmers, addressing the need for comprehensive data to understand cocoa cultivation [5]. In doing so, this study addresses important gaps that earlier investigations left open.

Additionally, this study builds upon previous findings, emphasizing the importance of weather conditions on evapotranspiration—a crucial factor for determining cocoa water requirements [31,49,50]. By recommending shade tree management practices, this research is consistent with prior studies [52] that support agroforestry systems as an effective strategy to mitigate temperature and moisture stress. This contribution is particularly significant given the climatic diversity in Colombia, which presents complex interactions between environmental variables and cocoa growth, as highlighted in the preliminary exploratory data analysis.

4.3. Limitations and Uncertainties

This methodological approach is particularly innovative because it integrates detailed datasets from sources like the NASA POWER database, which provides insights that address the previously noted absence of comprehensive environmental data [43,48]. While the computational setup employed an Intel Xeon CPU with 12.67 GB of RAM, the absence of GPU acceleration posed some limitations in scaling the more complex models, such as neural networks, to larger datasets. However, carefully selecting model parameters and preprocessing strategies mitigated these challenges, enabling efficient and accurate classification within the given computational constraints. For instance, decomposing k-means clusters into subsets facilitated scalability, ensuring smooth execution without exceeding memory limits. The results demonstrate that this configuration effectively handled the experiments' demands, supporting the reproducibility and accessibility of the methodology in resource-constrained environments. Despite the robustness of the best model, uncertainties such as data variability and localized weather events may impact the generalizability of the findings.

Relying on historical data and predefined models may only partially capture the dynamic nature of climate change, which continues to evolve rapidly. Changes such as increased droughts, floods, and storms introduce greater variability and exacerbate the existing vulnerabilities in agricultural systems [13]. While robust, the models used in this study are inherently limited by the resolution and scope of the input data, potentially overlooking localized environmental changes that could prove significant for cocoa cultivation. For instance, while cocoa is sensitive to drought stress, Colombia's high rainfall suggests that precipitation is not always a primary factor in determining land suitability. Yet, these environmental conditions vary greatly across global regions, making interpreting findings within a localized context necessary.

Another area for improvement is the potential under-representation of certain cocoa-growing regions that may need comprehensive historical data, especially in those farms with high area and variability across the land [55]. The variability across Colombia's agroecological zones requires more dynamic modeling to capture localized changes effectively. Future work should incorporate real-time environmental data to improve the models' adaptability and relevance to changing climatic conditions. The findings of this study contribute to the theoretical understanding of agroclimatic suitability for cocoa, demonstrating how advanced data-driven techniques can identify nuanced environmental influences on crop production. Future research should focus on developing more dynamic models incorporating real-time environmental data and machine-learning models capable of adapting to changing conditions. Progress in data collection and modeling could help address these regional disparities.

Another limitation is the overfitting risk; several robust strategies were implemented to mitigate that risk to ensure model generalization and reliability. The k-fold cross-validation

approach was employed to evaluate model performance across diverse subsets of data, effectively reducing overfitting risks and enhancing consistency. In the case of the ANN, a single hidden layer with 100 neurons was configured, utilizing the ReLU activation function to model non-linear relationships without introducing unnecessary complexity. Additionally, L2 regularization was applied within the ANN to penalize large weights, thereby preventing the model from becoming excessively complex and fitting noise in the data.

4.4. Future Directions

Future work will explore hyperparameter optimization strategies to enhance forecasting model performance while maintaining robust safeguards against overfitting. Techniques such as grid search, random search, and advanced methods like Bayesian optimization or evolutionary algorithms will be systematically tested to identify the optimal configurations for each model. These approaches must be complemented by extensive cross-validation and the use of validation curves to monitor overfitting risks, ensuring that the enhanced models maintain their predictive reliability across unseen data. A strategic blend of parameter tuning and rigorous validation can further elevate accuracy.

Lastly, this research contributes to the theoretical understanding of agroclimatic suitability for cocoa by demonstrating how advanced data-driven techniques can identify nuanced environmental influences on crop production. Applying clustering-based approaches to other crops or geographic regions may provide insights into the adaptability and scalability of these models in diverse agricultural contexts. Additionally, integrating advanced technologies, such as IoT sensors and satellite-based monitoring, could facilitate the implementation of these strategies, thereby enhancing productivity and sustainability in cocoa farming and supporting Colombia's position in the global cocoa market.

5. Conclusions

This study explored the key environmental factors affecting cocoa establishment in Colombia, focusing on water management and overall cultivation suitability. Given the increasing global demand for cocoa, understanding these environmental influences is crucial for improving productivity and ensuring sustainability in Colombia, a country with considerable potential for expanding its cocoa production. This research employed advanced machine learning techniques, including supervised and unsupervised models and ensemble methods, to analyze diverse environmental datasets and provide actionable insights into cocoa farming suitability under various conditions. The findings highlighted that temperature, humidity, and wind speed are critical determinants of cocoa growth, with specific interactions among these factors influencing the suitability of different regions for cocoa cultivation. Integrating crop growth models with hydrological data within an ensemble machine learning framework is innovative, offering practical recommendations for improving agricultural practices and enhancing resilience to climate change.

The model performance analysis revealed that random forest achieved the highest accuracy, effectively handling non-linear relationships and providing balanced metrics across suitability levels. Simpler models like logistic regression showed limited performance on the full dataset, but clustering significantly improved their accuracy within homogeneous clusters. This highlights the value of clustering in improving model interpretability and effectiveness by focusing on specific environmental conditions. At the same time, the ensemble approach provided nuanced, region-specific insights into cocoa cultivation suitability. Nevertheless, this study also encountered limitations, such as data resolution constraints and challenges in capturing localized climate dynamics. Looking ahead, the broader impact of this research lies in its potential to guide adaptive agricultural practices

and promote sustainable cocoa production, ultimately contributing to the resilience and growth of Colombia's agricultural sector.

Author Contributions: Conceptualization, L.T.-S.; methodology, L.T.-S.; software, L.T.-S.; validation, S.R.-P., L.C.-C. and O.G.-C.; formal analysis, L.T.-S.; investigation, L.T.-S.; resources, S.R.-P.; data curation, L.T.-S.; writing—original draft preparation, L.T.-S. and L.C.-C.; writing—review and editing, L.T.-S. and S.R.-P.; visualization, L.T.-S.; supervision, S.R.-P.; project administration, S.R.-P.; funding acquisition, S.R.-P. All authors have read and agreed to the published version of the manuscript.

Funding: The Colombian National Government supported this work through the Sistema General de Regalias (SRG) platform under the "Hallbar Kakao 2.0", code BPIN 2021000100362 project. The sponsors had no involvement in the study design, data collection, analysis, interpretation, writing, or decision to submit this article for publication.

Data Availability Statement: The data supporting this study's reported results are publicly available online at https://www.kaggle.com/datasets/lehetasa/colombian-cocoa-dataset under the CC BY-NC-SA 4.0 license. accessed on 1 October 2024. The datasets include environmental and topographical data from the NASA POWER database and agricultural data from official Colombian agricultural sources.

Acknowledgments: The authors thank the administrative and technical staff at the Universidad Autonoma de Bucaramanga for their invaluable support throughout the project. They also thank the Hallbar Kakao 2.0 project researchers for their insightful contributions and Mauren Slendy Cardenas Fontecha and Leivis Milagro Pua de la Hoz for their moral support and administrative guidance. They also thank the Direccion de Investigaciones at the Universidad Autonoma de Bucaramanga for their support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Díaz-Valderrama, J.R.; Leiva-Espinoza, S.T.; Aime, M.C. The history of cacao and its diseases in the Americas. *Phytopathology* **2020**, *110*, 1604–1619. [CrossRef] [PubMed]
2. Burgon, V.H.; Silva, M.L.N.; Milani, R.F.; Morgano, M.A. Trace elements in bean-to-bar chocolates from Brazil and Ecuador. *J. Trace Elem. Med. Biol.* **2024**, *84*, 127431. [CrossRef]
3. AGROSAVIA. El Cacao Una Historia Que Se Está Escribiendo. 2023. Available online: https://www.agrosavia.co/noticias/el-cacao-una-historia-que-se-est%C3%A1-escribiendo (accessed on 1 September 2024).
4. Kongor, J.E.; Hinneh, M.; Van de Walle, D.; Afoakwa, E.O.; Boeckx, P.; Dewettinck, K. Factors influencing quality variation in cocoa (Theobroma cacao) bean flavour profile-A review. *Food Res. Int.* **2016**, *82*, 44–52. [CrossRef]
5. Fernández-Niño, M.; Rodríguez-Cubillos, M.J.; Herrera-Rocha, F.; Anzola, J.M.; Cepeda-Hernández, M.L.; Mejía, J.L.A.; Chica, M.J.; Olarte, H.H.; Rodríguez-López, C.; Calderón, D.; et al. Dissecting industrial fermentations of fine flavour cocoa through metagenomic analysis. *Sci. Rep.* **2021**, *11*, 8638. [CrossRef] [PubMed]
6. Escobar, S.; Santander, M.; Zuluaga, M.; Chacón, I.; Rodríguez, J.; Vaillant, F. Fine cocoa beans production: Tracking aroma precursors through a comprehensive analysis of flavor attributes formation. *Food Chem.* **2021**, *365*, 130627. [CrossRef] [PubMed]
7. Bacca-Villota, P.; Acuña-García, L.; Sierra-Guevara, L.; Cano, H.; Hidalgo, W. Untargeted Metabolomics Analysis for Studying Differences in High-Quality Colombian Cocoa Beans. *Molecules* **2023**, *28*, 4467. [CrossRef] [PubMed]
8. Escobar, S.; Santander, M.; Useche, P.; Contreras, C.; Rodríguez, J. Aligning strategic objectives with research and development activities in a soft commodity sector: A technological plan for colombian cocoa producers. *Agriculture* **2020**, *10*, 141. [CrossRef]
9. Pemberton, C.A.; De Sormeaux, A.; Patterson-Andrews, H. Comparing the volatility of the international prices of cocoa, coffee and oil. *Trop. Agric.* **2018**, *95*, 181–193. Available online: https://journals.sta.uwi.edu/ojs/index.php/ta/article/view/6627 (accessed on 1 September 2024).
10. Diaz, R.T.; Osorio, D.P.; Hernández, E.M.; Pallares, M.M.; Canales, F.A.; Paternina, A.C.; Echeverría-González, A. Socioeconomic determinants that influence the agricultural practices of small farm families in northern Colombia. *J. Saudi Soc. Agric. Sci.* **2022**, *21*, 440–451. [CrossRef]
11. Departamento Nacional de Planeación. *El Campo Colombiano: Un Camino Hacia el Bienestar y la Paz*; DNP: Bogotá, Colombia, 2015.
12. Parra-Paitan, C.; Meyfroidt, P.; Verburg, P.H.; Ermgassen, E.K.H.J.Z. Deforestation and climate risk hotspots in the global cocoa value chain. *Environ. Sci. Policy* **2024**, *158*, 103796. [CrossRef]

13. Parry, M.L. *Climate Change and World Agriculture*; Taylor & Francis: London, UK, 2019. [CrossRef]
14. Akrofi, A.Y.; Amoako-Atta, I.; Assuah, M.; Asare, E.K. Black pod disease on cacao (*Theobroma cacao* L.) in Ghana: Spread of Phytophthora megakarya and role of economic plants in the disease epidemiology. *Crop Prot.* **2015**, *72*, 66–75. [CrossRef]
15. Mpika, J.; Kebe, I.B.; N'Guessan, K.F. Isolation and Identification of Indigenous Microorganisms of Cocoa Farms in Côte d'Ivoire and Assessment of Their Antagonistic Effects Vis-À-Vis Phytophthora palmivora, the Causal Agent of the Black Pod Disease. *Biodivers. Loss A Change Planet* **2011**, *11*, 13. [CrossRef]
16. Opoku, I.; Appiah, A.; Akrofi, A.; Owusu, G. Phytophthora megakarya: A potential threat to the cocoa industry in Ghana. *Ghana J. Agric. Sci.* **2000**, *33*, 237–248. [CrossRef]
17. Mahrizal; Nalley, L.L.; Dixon, B.L.; Popp, J.S. An optimal phased replanting approach for cocoa trees with application to Ghana. *Agric. Econ.* **2014**, *45*, 291–302. [CrossRef]
18. Adomako, B.; Adu-Ampomah, Y. Reflections on the yield of Upper Amazon cocoa hybrids in Ghana with reference to breeding for cocoa swollen shoot virus resistant varieties. *Cocoa Grow. Bull.* **2000**, *52*, 33–45.
19. Edwin, J.; Masters, W.A. Genetic improvement and cocoa yields in ghana. *Exp Agric* **2005**, *41*, 491–503. [CrossRef]
20. Appiah, M.R.; Ofori-Frimpong, K.; Afrifa, A. Evaluation of fertilizer application on some peasant cocoa farms in Ghana. *Ghana J. Agric. Sci.* **2000**, *33*, 183–190. [CrossRef]
21. Baah, F.; Anchirinah, V.; Amon-Armah, F. Soil fertility management practices of cocoa farmers in the Eastern Region of Ghana. *Agric. Biol. J. N. Am.* **2011**, *2*, 173–181. [CrossRef]
22. Sonwa, D.J.; Weise, S.F.; Schroth, G.; Janssens, M.J.J.; Shapiro, H.-Y. Structure of cocoa farming systems in West and Central Africa: A review. *Agrofor. Syst.* **2019**, *93*, 2009–2025. [CrossRef]
23. Souza, C.A.S.; Dias, L.A.d.S.; Aguilar, M.A.G.; Sonegheti, S.; Oliveira, J.; Costa, J.L.A. Cacao yield in different planting densities. *Braz. Arch. Biol. Technol.* **2009**, *52*, 1313–1320. [CrossRef]
24. Schroth, G.; Läderach, P.; Martinez-Valle, A.I.; Bunn, C.; Jassogne, L. Vulnerability to climate change of cocoa in West Africa: Patterns, opportunities and limits to adaptation. *Sci. Total Environ.* **2016**, *556*, 231–241. [CrossRef] [PubMed]
25. Mensah, E.O.; Vaast, P.; Asare, R.; Amoatey, C.A.; Owusu, K.; Asitoakor, B.K.; Ræbild, A. Cocoa Under Heat and Drought Stress. In *Agroforestry as Climate Change Adaptation*; Springer International Publishing: Cham, Switzerland, 2024; pp. 35–57. [CrossRef]
26. Kongor, J.E.; Owusu, M.; Oduro-Yeboah, C. Cocoa production in the 2020s: Challenges and solutions. *CABI Agric. Biosci.* **2024**, *5*, 102. [CrossRef]
27. Igawa, T.K.; de Toledo, P.M.; Anjos, L.J.S. Climate change could reduce and spatially reconfigure cocoa cultivation in the Brazilian Amazon by 2050. *PLoS ONE* **2022**, *17*, e0262729. [CrossRef]
28. Läderach, P.; Martinez-Valle, A.; Schroth, G.; Castro, N. Predicting the future climatic suitability for cocoa farming of the world's leading producer countries, Ghana and Côte d'Ivoire. *Clim. Change* **2013**, *119*, 841–854. [CrossRef]
29. Andreotti, F.; Mao, Z.; Jagoret, P.; Speelman, E.N.; Gary, C.; Saj, S. Exploring management strategies to enhance the provision of ecosystem services in complex smallholder agroforestry systems. *Ecol. Indic.* **2018**, *94*, 257–265. [CrossRef]
30. Niether, W.; Schneidewind, U.; Fuchs, M.; Schneider, M.; Armengot, L. Below- and aboveground production in cocoa monocultures and agroforestry systems. *Sci. Total Environ.* **2019**, *657*, 558–567. [CrossRef]
31. Acheampong, K.; Daymond, A.J.; Adu-Yeboah, P.; Hadley, P. Improving field establishment of cacao (Theobroma cacao) through mulching, irrigation and shading. *Exp. Agric.* **2019**, *55*, 898–912. [CrossRef]
32. Ofori-Boateng, K.; Insah, B. The impact of climate change on cocoa production in West Africa. *Int. J. Clim. Change Strat. Manag.* **2014**, *6*, 296–314. [CrossRef]
33. Ntiamoah, A.; Afrane, G. Environmental impacts of cocoa production and processing in Ghana: Life cycle assessment approach. *J. Clean. Prod.* **2008**, *16*, 1735–1740. [CrossRef]
34. Lamos-Díaz, H.; Puentes-Garzón, D.E.; Zarate-Caicedo, D.A. Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia. *Rev. Fac. De Ing.* **2020**, *29*, e10853. [CrossRef]
35. Talero-Sarmiento, L.H.; Parra-Sanchez, D.T.; Diaz, H.L. Opportunities and Barriers of Smart Farming Adoption by Farmers Based on a Systematic Literature Review. In Proceedings of the INNODOCT/22, International Conference on Innovation, Documentation and Education, Valencia, FL, USA, 2–7 November 2023; pp. 53–64. [CrossRef]
36. Hajjaji, Y.; Boulila, W.; Farah, I.R.; Romdhani, I.; Hussain, A. Big data and IoT-based applications in smart environments: A systematic review. *Comput. Sci. Rev.* **2021**, *39*, 100318. [CrossRef]
37. Talero-Sarmiento, L.H.; Parra-Sanchez, D.T.; Lamos-Diaz, H. A Bibliometric Analysis of Computational and Mathematical Techniques in the Cocoa Sustainable Food Value Chain. *arXiv* **2023**, 1–61. [CrossRef]
38. Araújo, S.O.; Peres, R.S.; Ramalho, J.C.; Lidon, F.; Barata, J. Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives. *Agronomy* **2023**, *13*, 2976. [CrossRef]
39. Tosto, A.; Morales, A.; Rahn, E.; Evers, J.B.; Zuidema, P.A.; Anten, N.P.R. Simulating cocoa production: A review of modelling approaches and gaps. *Agric. Syst.* **2023**, *206*, 103614. [CrossRef]

40. Meshram, V.; Patil, K.; Meshram, V.; Hanchate, D.; Ramkteke, S.D. Machine learning in agriculture domain: A state-of-art survey. *Artif. Intell. Life Sci.* **2021**, *1*, 100010. [CrossRef]
41. Gawdiya, S.; Kumar, D.; Ahmed, B.; Sharma, R.K.; Das, P.; Choudhary, M.; Mattar, M.A. Field scale wheat yield prediction using ensemble machine learning techniques. *Smart Agric. Technol.* **2024**, *9*, 100543. [CrossRef]
42. Cravero, A.; Pardo, S.; Sepúlveda, S.; Muñoz, L. Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review. *Agronomy* **2022**, *12*, 748. [CrossRef]
43. Condran, S.; Bewong, M.; Islam, M.Z.; Maphosa, L.; Zheng, L. Machine Learning in Precision Agriculture: A Survey on Trends, Applications and Evaluations over Two Decades. *IEEE Access* **2022**, *10*, 73786–73803. [CrossRef]
44. Ayed, R.B.; Hanana, M. Artificial Intelligence to Improve the Food and Agriculture Sector. *J. Food Qual.* **2021**, *2021*, 5584754. [CrossRef]
45. Heredia-Gómez, J.F.; Rueda-Gómez, J.P.; Talero-Sarmiento, L.H.; Ramírez-Acuña, J.S.; Coronado-Silva, R.A. Cocoa pods ripeness estimation, using convolutional neural networks in an embedded system. *Rev. Colomb. Comput.* **2020**, *21*, 42–55. [CrossRef]
46. Arenga, D.Z.H.; Cruz, J.C.D. Ripeness classification of cocoa through acoustic sensing and machine learning. In Proceedings of the HNICEM 2017-9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, Manila, Philippines, 1–3 December 2017. [CrossRef]
47. De Oliveira, J.R.C.P.; Romero, R.A.F. Transfer Learning Based Model for Classification of Cocoa Pods. In Proceedings of the International Joint Conference on Neural Networks, Rio de Janeiro, Brazil, 8–13 July 2018. [CrossRef]
48. Gamboa, A.A.; Caceres, P.A.; Lamos, H.; Zarate, D.A.; Puentes, D.E. Predictive model for cocoa yield in Santander using Supervised Machine Learning. In Proceedings of the 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), Bucaramanga, Colombia, 24–26 April 2019; pp. 1–5. [CrossRef]
49. Jezeer, R.E.; Verweij, P.A.; Santos, M.J.; Boot, R.G.A. Shaded Coffee and Cocoa–Double Dividend for Biodiversity and Small-scale Farmers. *Ecol. Econ.* **2017**, *140*, 136–145. [CrossRef]
50. Niether, W.; Armengot, L.; Andres, C.; Schneider, M.; Gerold, G. Shade trees and tree pruning alter throughfall and microclimate in cocoa (*Theobroma cacao* L.) production systems. *Ann. Sci.* **2018**, *75*, 38. [CrossRef]
51. Akpalu, M.M.; Ofosu-Budu, G.K.; Kumaga, F.K.; Ofori, K.; Oppong-Danso, E. Mulching and Irrigation Practices on Cocoa Seedling Survival and Field Establishment. *J. Agric. Crops* **2020**, *6*, 126–132. [CrossRef]
52. Cilas, C.; Bastide, P. Challenges to Cocoa Production in the Face of Climate Change and the Spread of Pests and Diseases. *Agronomy* **2020**, *10*, 1232. [CrossRef]
53. Niether, W.; Jacobi, J.; Blaser, W.J.; Andres, C.; Armengot, L. Cocoa agroforestry systems versus monocultures: A multi-dimensional meta-analysis. *Environ. Res. Lett.* **2020**, *15*, 104085. [CrossRef]
54. Unidad de Planificación Rural Agropecuaria-UPRA. Zonificación de Aptitud para el Cultivo Comercial de Cacao (*Theobroma cacao* L.) en Colombia. Bogotá, D.C., Colombia. 2024. Available online: https://sipra.upra.gov.co (accessed on 1 September 2024).
55. Brook, A.; De Micco, V.; Battipaglia, G.; Erbaggio, A.; Ludeno, G.; Catapano, I.; Bonfante, A. A smart multiple spatial and temporal resolution system to support precision agriculture from satellite images: Proof of concept on Aglianico vineyard. *Remote Sens. Environ.* **2020**, *240*, 111679. [CrossRef]
56. Letu, H.; Shi, J.; Li, M.; Wang, T.; Shang, H.; Lei, Y.; Ji, D.; Wen, J.; Yang, K.; Chen, L. A review of the estimation of downward surface shortwave radiation based on satellite data: Methods, progress and problems. *Sci. China Earth Sci.* **2020**, *63*, 774–789. [CrossRef]
57. Wang, F.; Tian, D.; Carroll, M. Customized deep learning for precipitation bias correction and downscaling. *Geosci. Model. Dev.* **2023**, *16*, 535–556. [CrossRef]
58. Nyamsi, W.W.; Espinar, B.; Blanc, P.; Wald, L. Estimating the photosynthetically active radiation under clear skies by means of a new approach. *Adv. Sci. Res.* **2015**, *12*, 5–10. [CrossRef]
59. Saavedra, F.; Peña, E.J.; Schneider, M.; Naoki, K. Effects of environmental variables and foliar traits on the transpiration rate of cocoa (*Theobroma cacao* L.) under different cultivation systems. *Agrofor. Syst.* **2020**, *94*, 2021–2031. [CrossRef]
60. Liuzzo, L.; Viola, F.; Noto, L.V. Wind speed and temperature trends impacts on reference evapotranspiration in Southern Italy. *Theor. Appl. Clim.* **2016**, *123*, 43–62. [CrossRef]
61. NASA, The Power Project. NASA Prediction Of Worldwide Energy Resources. 2022. Available online: https://power.larc.nasa.gov/ (accessed on 22 August 2022).
62. Sperandei, S. Understanding logistic regression analysis. *Biochem. Med.* **2014**, *24*, 12–18. [CrossRef] [PubMed]
63. Safavian, S.R.; Landgrebe, D. A Survey of Decision Tree Classifier Methodology. *IEEE Trans. Syst. Man. Cybern.* **1991**, *21*, 660–674. [CrossRef]
64. Genuer, R.; Poggi, J.-M. *Random Forests with R*; Springer: Dordrecht, The Netherlands, 2020. [CrossRef]
65. Chang, C.C.; Lin, C.J. LIBSVM: A Library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–39. [CrossRef]
66. Windeatt, T. Ensemble MLP classifier design. In *Studies in Computational Intelligence*; Springer: Berlin/Heidelberg, Germany, 2008; p. 137. [CrossRef]

67. Sinaga, K.P.; Yang, M.S. Unsupervised K-means clustering algorithm. *IEEE Access* **2020**, *8*, 80716–80727. [CrossRef]
68. Ledoit, O.; Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* **2004**, *88*, 365–411. [CrossRef]
69. Pedregosa, F. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
70. Hunter, J.; Dale, D.; Firing, E.; Droettboom, M. Matplotlib: Visualization with Python. *Abgerufen* **2020**.
71. Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 92–96. [CrossRef]
72. Kongor, J.E.; Boeckx, P.; Vermeir, P.; Van de Walle, D.; Baert, G.; Afoakwa, E.O.; Dewettinck, K. Assessment of soil fertility and quality for improved cocoa production in six cocoa growing regions in Ghana. *Agrofor. Syst.* **2019**, *93*, 1455–1467. [CrossRef]
73. Essougong, U.P.K.; Slingerland, M.; Mathé, S.; Vanhove, W.; Ngome, P.I.T.; Boudes, P.; Giller, K.E.; Woittiez, L.S.; Leeuwis, C. Farmers' Perceptions as a Driver of Agricultural Practices: Understanding Soil Fertility Management Practices in Cocoa Agroforestry Systems in Cameroon. *Hum. Ecol.* **2020**, *48*, 709–720. [CrossRef]
74. Daymond, A.; Mendez, D.G.; Hadley, P.; Bastide, P. *A Global Review of Cocoa Farming Systems*; Whiteknights: Reading, UK; Montpellier, France, 2021. Available online: https://www.icco.org/wp-content/uploads/Global-Review-of-Cocoa-Farming-Systems_Final.pdf (accessed on 1 September 2024).
75. Asigbaase, M.; Lomax, B.H.; Dawoe, E.; Sjogersten, S. Influence of organic cocoa agroforestry on soil physico-chemical properties and crop yields of smallholders' cocoa farms, Ghana. *Renew. Agric. Food Syst.* **2020**, *36*, 255–264. [CrossRef]
76. Nahon, S.M.R.; Trindade, F.C.; Yoshiura, C.A.; Martins, G.C.; da Costa, I.R.C.; Costa, P.H.d.O.; Herrera, H.; Balestrin, D.; Godinho, T.d.O.; Marchiori, B.M.; et al. Impact of Agroforestry Practices on Soil Microbial Diversity and Nutrient Cycling in Atlantic Rainforest Cocoa Systems. *Int. J. Mol. Sci.* **2024**, *25*, 11345. [CrossRef] [PubMed]
77. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
78. Tetko, I.V.; Livingstone, D.J.; Luik, A.I. Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826–833. [CrossRef]
79. Razaque, A.; Frej, M.B.H.; Almi'ani, M.; Alotaibi, M.; Alotaibi, B. Improved Support Vector Machine Enabled Radial Basis Function and Linear Variants for Remote Sensing Image Classification. *Sensors* **2021**, *21*, 4431. [CrossRef] [PubMed]
80. Reis, G.S.M.; de Almeida, A.A.F.; Mangabeira, P.A.O.; Santos, I.C.D.; Pirovani, C.P.; Ahnert, D. Mechanical stress caused by wind on leaves of Theobroma cacao: Photosynthetic, molecular, antioxidative and ultrastructural responses. *PLoS ONE* **2018**, *13*, e0198274. [CrossRef] [PubMed]