agriculture

MDPI

Article

# Effectiveness of Common Preprocessing Methods of Time Series for Monitoring Crop Distribution in Kenya

Rui Ni [1,2], Xiaohui Zhu [3], Yuping Lei [1], Xiaoxin Li [1], Wenxu Dong [1], Chuang Zhang [1,2], Tuo Chen [1], David M. Mburu [4] and Chunsheng Hu [1,2,*]

[1] Center for Agricultural Resources Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Hebei Laboratory of Agricultural Water-Saving, Key Laboratory of Agricultural Water Resources, Shijiazhuang 050022, China; nirui19@mails.ucas.ac.cn (R.N.); leiyp@sjziam.ac.cn (Y.L.); xiaoxin_li@sjziam.ac.cn (X.L.); dongwx@sjziam.ac.cn (W.D.); zhcnd10@126.com (C.Z.); tuochen@sjziam.ac.cn (T.C.)

[2] College of Advanced Agricultural Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[3] Department of Earth and Environment, Boston University, Boston, MA 02215, USA; zhuxh@bu.edu

[4] College of Agriculture and Natural Resources, Jomo Kenyatta University of Agriculture and Technology, Nairobi P.O. Box 62000-00200, Kenya; dmburu.mburu13@gmail.com

* Correspondence: cshu@sjziam.ac.cn

**Abstract:** Accurate crop identification and spatial distribution mapping are important for crop production estimation and famine early warning, especially for food-deficit African agricultural countries. By evaluating existing preprocessing methods for classification using satellite image time series (SITS) in Kenya, this study aimed to provide a low-cost method for cultivated land monitoring in sub-Saharan Africa that lacks financial support. SITS were composed of a set of MODIS Vegetation Indices (MOD13Q1) in 2018, and the classification method included the Support Vector Machine (SVM) and Random Forest (RF) classifier. Eight datasets obtained at three levels of preprocessing from MOD13Q1 were used in the classification: (1) raw SITS of vegetation indices (R-NDVI, R-EVI, and R-NDVI + R-EVI); (2) smoothed SITS of vegetation indices (S-NDVI); and (3) vegetation phenological data (P-NDVI, P-EVI, R-NDVI + P-NDVI, and P-NDVI-1). Both SVM and RF classification results showed that the "R-NDVI + R-EVI" dataset achieved the highest performance, while the three pure phenological datasets produced the lowest accuracy. Correlation analysis between variable importance and rainfall time series demonstrated that the vegetation index SITS during rainfall periods showed higher importance in RF classifiers, thus revealing the potential of saving computational costs. Considering the preprocessing cost of SITS and its negative impact on the classification accuracy, we recommend overlaying the original NDVI with the original EVI time series to map the crop distribution in Kenya.

**Keywords:** Kenya; satellite image time series; MODIS; random forest; support vector machine; cropland; TIMESAT; phenometrics

check for updates

**Citation:** Ni, R.; Zhu, X.; Lei, Y.; Li, X.; Dong, W.; Zhang, C.; Chen, T.; Mburu, D.M.; Hu, C. Effectiveness of Common Preprocessing Methods of Time Series for Monitoring Crop Distribution in Kenya. *Agriculture* **2022**, *12*, 79. https://doi.org/10.3390/agriculture12010079

Academic Editor: Koki Homma

Received: 22 December 2021
Accepted: 5 January 2022
Published: 7 January 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## 1. Introduction

Agriculture is one of the most important pillar industries of Kenya. According to the World Bank, agricultural output accounted for 31% of the country's GDP in 2014. However, less than 20% of the land is suitable for agricultural production, of which only 12% is of high agricultural potential (due to adequate rainfall), and about 8% is of medium agricultural potential [1]. At the same time, Kenya is one of the countries most affected by hunger in the world due to the severe lack of agricultural inputs and the impact of natural disasters. According to the Global Hunger Index 2020 published by the International Food Policy Research Institute (IFPRI), Kenya is ranked 84th out of 107 countries in terms of food security, with hunger levels defined as "serious." Therefore, timely and reliable mapping of crop cultivation is essential for developing sound food policies and responding quickly to emerging food shortages [2]. With the rapid development of spatial information technology, satellite imagery has been widely adopted in the fields of crop identification mapping and growth monitoring. Its characteristics of wide coverage, timeliness, low investment, and high return make it a reliable approach for accurately and rapidly monitoring cultivated areas and crop growth and supporting the implementation of agricultural policies.

The typical cropping status and climatic characteristics of the Kenyan region are major obstacles in crop distribution monitoring. Smallholder farms are the most common form of agricultural production in sub-Saharan African countries, e.g., Kenya [3]. For example, in an analysis of subnational census data, Samberg et al. [4] reported that 50% of food calories in sub-Saharan Africa come from farms less than 5 ha in size. A major challenge of land cover mapping in the tropics is the high degree of heterogeneity. It means many mixed pixels containing multiple agricultural cover types dominated the images of the region [5]. Therefore, medium- and high-resolution images have become common data sources in subnational studies for land cover and crop distribution mapping [6–9]. Notably, several aspects of smallholder farming in the tropics intensified the challenges for identifying crops using remote sensing. Firstly, Maingi and Marsh [10] and Tottrup [11] in their respective studies indicated that the use of multispectral images acquired by satellites (such as Landsat) showed less spectral separation among vegetation types in tropical forest areas. Secondly, cloud-free images are difficult to achieve in the area during the crop growing season, that is, the rainy season. According to Hashim et al. [12], 75% of the satellite images acquired over the equator are obscured by clouds. Moreover, in the Kenyan region, the growing season mostly overlaps with the rainy season, and each vegetation type rapidly enters the developmental phase whenever the rainy season arrives. The frequent cloudy and rapid response of crops to the rainy season pushes the demands for higher temporal resolution of remote sensing imagery. Due to the trade-off between spatial and temporal resolutions, satellites with higher spatial resolution (e.g., Landsat, SPOT, and Sentinel) cannot meet the temporal resolution requirement of classification, specifically reflected in the high cloud coverage of images obtained during the growing season (i.e., rainy season). Finally, the use of high spatial and temporal resolution imagery in large-scale areas (e.g., a national scale) requires high image costs' and processing costs' input, which always constrain the application of crop mapping for most African countries that lack financial support.

Moderate Resolution Imaging Spectroradiometer (MODIS) time series has become a common classification strategy for crop distribution mapping in tropical high-cloud areas. Vithanage et al. [13] suggested that spectral variation throughout the year can be a significant feature for tropical and subtropical regions dominated by agriculture, and standard classification methods cannot account for temporal or spectral variation across multiple seasons. For land cover mapping in the Rift Valley Province of Kenya, Baldyga et al. [14] indicated the constraint of strong temporal variability on land cover mapping and the importance of the time series of vegetation indices for describing this variability. Vegetation index profiles, driven by vegetation phenology and influenced by differences in land conditions, climatic conditions, and cropping habits, reflect unique vegetation characteristics that can be exploited [15], and the vegetation index processing signal varies seasonally with growth patterns. Therefore, satellite image time series (SITS) of vegetation indices are commonly used to map crop distribution in Kenya [1,13,16–20]. Generally, SITS monitors the regional vegetation dynamics at the shortest possible time interval, thereby addressing the problem of high cloud cover.

To map crop distribution using SITS, many studies have been conducted using different input datasets with various classification methods. Some of these studies have utilized smoothed preprocessed vegetation index time series (e.g., NDVI or EVI) as the dataset involved in the classification [21–24]. The purpose of using smoothing algorithms is to eliminate noise levels in the original datasets, especially in tropical and subtropical regions with substantial cloud cover, so as to provide a more reliable time-series dataset. While other studies have focused on exploring the extraction of vegetation phenometrics (e.g., growing season start/end, growing season length) or statistical values (e.g., mean, extreme values, amplitude, median, and standard deviation) from SITS, and then used as the basis for crop identification. [1,18,25]. For example, Valero et al. [26] found that the statistical values extracted from the NDVI profile showed their ability to improve the characterization of some land cover categories. Chen et al. [27] calculated six phenometrics (i.e., start of season, end of season, maximum value, amplitude, base value, and length of growing season) from SITS of MODIS NDVI to identify crop types and cropping patterns in the state of Mato Grosso, Brazil. Pelletier et al. [28] used the extracted nine phenometrics (including the beginning, the peak, and the length of the season) as part of the input features to evaluate the robustness of the machine learning classifier for accurate land cover mapping over large areas. Nonetheless, Kuchler et al. [29] and Araujo Picoli et al. [30] questioned the feasibility of noise cancellation and curve smoothing in mapping cropping systems and land cover classifications, arguing that it might reduce the amount of information in the time series of vegetation indices. Jamel et al. [31] also demonstrated the ability of advanced classification algorithms to handle noisy high-dimensional data and, hence, advocated for the classification of origin data when using statistical machine learning algorithms.

Since the usual preprocessing methods of SITS (time series denoising and phenometrics′ extraction) are time-consuming and laborious, it is worth exploring whether such preprocessing has a positive or negative impact on the classification results, especially when machine learning methods were used. Therefore, this work proposed to build an affordable method for agricultural mapping in sub-Saharan Africa that lacks financial support by evaluating the accuracy of common preprocessing methods of MODIS time series for crop identification in Kenya. For this purpose, we derived eight datasets with different preprocessing levels from MOD13Q1 time series data and compared the accuracies of the classification of various types of land cover in the Kenyan region in 2018, using a widely used machine learning classification algorithm (i.e., support vector machine and random forest). We hypothesized that land cover classification using the random forest algorithm for the full depth of the original time series is a robust and effective classification method for the Kenyan region.

## 2. Materials and Methods

### *2.1. Study Area*

The Republic of Kenya (04°40′ S–05°02′ N and 33°56′ E–41°34′ E) is located in the eastern part of Africa, where the equator crosses the central part, and has an area of 582,646 km$^2$. The southeast is bordered by the Indian Ocean, with a coastline of 536 km. The year-round highest temperature ranges between 22–26 °C while the lowest temperature ranges between 10–14 °C. While the whole territory is located in the tropical monsoon region, it has a savanna climate due to higher terrain. It has significant seasonal differences in precipitation, with a long rainy season from March to May, a short rainy season from October to December, and a dry season for the rest of the year. According to the rainfall distribution in Kenya, only 16% of the country receives more than 750 mm of annual rainfall [32]. The arable land area is about 105,000 km$^2$, mainly distributed in the areas around Mount Elgon National Park to Aberdare Park–Mount Kenya and Lake Victoria in the west [33].

As one of the largest maize producers in sub-Saharan Africa, Kenya has a typical maize cropping pattern of smallholder fields intercropped with other major crops (e.g., beans, cowpeas, sorghum, and millet) or cash crops. In addition to maize, Kenya also grows wheat, sugarcane, tea, rice, coffee, and watermelon. In this study, we focused on extracting the distribution of major crops throughout Kenya to achieve an accurate and rapid estimation of crop acreage.

### *2.2. Data*

Given the requirements of swath width (by the scale of the study area) and the requirements of temporal resolution (by meteorological conditions), we selected Terra MODIS Vegetation Indices (MOD13Q1) as the primary imagery data source in this study.

With a spatial resolution of 250 m and a swath width of 1200 km, MOD13Q1 can cover entire Kenya in four images. Through the maximum value composite (MVC), MOD13Q1 can provide the best available pixel values of daily acquisition for every 16-day period and tackle the issue of cloud pollution to a certain extent. The complete SITS of MOD13Q1 for 2000, 2015, 2017, 2018, 2019, and 2020 were downloaded from the Google Earth Engine (GEE) platform. The acquisition, mosaicking, cutting, and reprojection of MODIS time series data covering the entire territory of Kenya were completed through a JavaScript interface. Two primary vegetation indices, Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) from the MOD13Q1 product, were adopted in this study. With the spectral transformation of the visible red band (strongly absorbed by chlorophyll) and the near-infrared band (high reflection and high transmission of leaf structure), NDVI can serve as an excellent indicator of vegetation growth state and vegetation coverage. EVI introduces canopy background adjustment, blue band, and atmosphere resistance to reduce the soil background and residual atmospheric pollution. Compared with NDVI, EVI is more sensitive to high biomass regions [34]. Eight datasets corresponding to different preprocessing levels were considered to evaluate the effects of the preprocessing for the vegetation index time series on crop distribution mapping in Kenya:

- R-NDVI: raw NDVI time series (23 bands, 23 NDVI image sequences per year).
- R-EVI: raw EVI time series (23 bands, 23 EVI image sequences per year).
- S-NDVI: smoothed NDVI time series (23 bands, 23 NDVI image sequences by smoothing).
- P-NDVI: phenological parameters obtained from the original NDVI time series (26 bands, i.e., 26 vegetation phenometrics extracted from two growing seasons).
- P-EVI: phenological parameters obtained from the original EVI time series (26 bands, vegetation phenometrics extracted from two growing seasons).
- P-NDVI-1: phenological parameters obtained from the original NDVI time series (13 bands, i.e., 13 vegetation phenometrics extracted from the first growing season).
- R-NDVI + R-EVI: a combination of original NDVI and original EVI time series (46 bands, 23 NDVI + 23 EVI).
- R-NDVI + P-NDVI: a combination of original NDVI time series and NDVI-derived phenological parameters (49 bands, 23 NDVI + 26 vegetation phenometrics).

The Kenya Crop Mask data for 2000 and 2015 were obtained from the Regional Centre for Mapping of Resources for Development (RCMRD) open dataset, which provides the extent of cropland, area-specific major crops, and other crops being grown at the same location. The layers were generated using Landsat 5 TM and Landsat 8 OLI in September/October 2000 and 2015, respectively, and validated by IN SITU data. Crop regions with no change in cropping structure during the 15 years (Figure 1a) were extracted by crop mask files. Based on this, the sample selection for crop identification was performed by the local experienced remote sensing interpretation experts in combination with years of very high-resolution satellite images from Google Earth. According to the main land cover and crop types in Kenya, the classification was defined as: (1) artificial surface; (2) bare land; (3) coffee; (4) forest and wetland; (5) herb and shrubland; (6) maize; (7) pineapple; (8) rice; (9) sisal; (10) sugarcane; (11) tea; (12) water body; (13) watermelon; and (14) wheat. It should be noted that, as the most widely grown and distributed crop in Kenya, maize is often mixed with legumes and other crops. Limited by the spatial resolution of the images, mixed legumes, sorghum, and small amounts of sporadically grown sugarcane, millet, vegetables, and fruits were ignored in the maize-growing areas in the classification in this study.

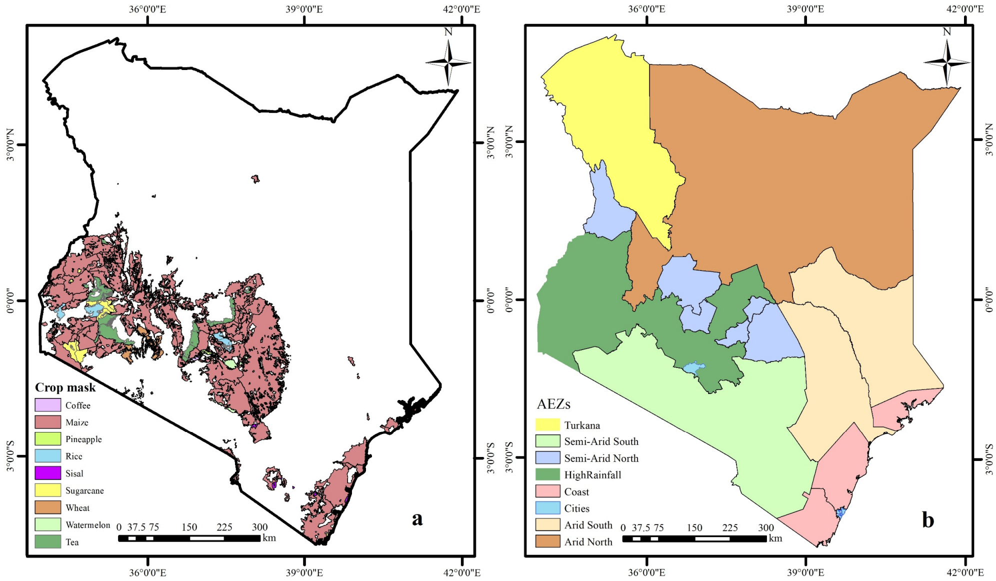

**Figure 1.** (**a**) Unchanged crop masks in Kenya between 2000 and 2015; (**b**) Kenya agroecological zones (AEZs).

Precipitation data were obtained from CHIRPS Daily (Climate Hazards Group InfraRed Precipitation with Station data) to analyze the correlations between the importance of classifier variables and rainfall. As a quasi-global daily precipitation dataset, CHIRPS combines 0.05° resolution satellite imagery and in situ station data to form a gridded precipitation time series for trend analysis and seasonal drought monitoring [35].

### 2.3. Methods

Here the general approach of the classification is displayed in Figure 2. To identify the best dataset for extracting crop distribution information in Kenya, eight datasets containing raw time series of vegetation indices (R-NDVI, R-EVI, and R-NDVI + R-EVI), smoothed time series of vegetation indices (S-NDVI), and vegetation phenology data (P-NDVI, P-EVI, and P-NDVI-1) were used. Random Forests' algorithms were used to generate crop spatial distribution maps based on the eight datasets, respectively.

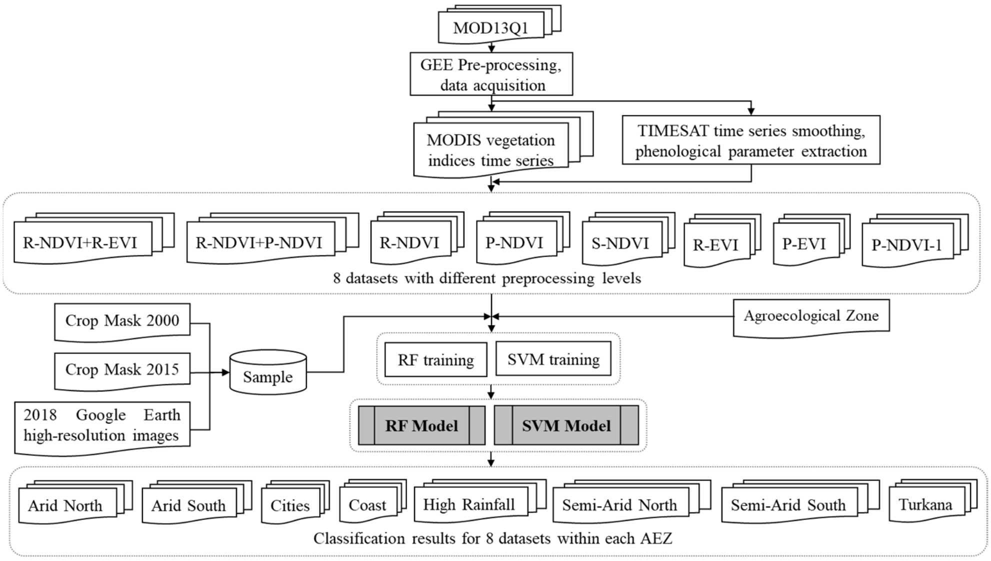
MOD13Q1
GEE Pre-processing, data acquisition
MODIS vegetation indices time series
TIMESAT time series smoothing, phenological parameter extraction
R-NDVI+R-EVI
R-NDVI+P-NDVI
R-NDVI
P-NDVI
S-NDVI
R-EVI
P-EVI
P-NDVI-1
8 datasets with different preprocessing levels
Crop Mask 2000
Crop Mask 2015
2018 Google Earth high-resolution images
Sample
Agroecological Zone
RF training
SVM training
RF Model
SVM Model
Arid North
Arid South
Cities
Coast
High Rainfall
Semi-Arid North
Semi-Arid South
Turkana
Classification results for 8 datasets within each AEZ

**Figure 2.** Graphical scheme of classification process in this study.

### 2.3.1. Vegetation Index Time Series Smoothing and Phenometrics' Extraction

The S-NDVI and P-NDVI/EVI datasets were obtained from original NDVI or EVI datasets after image smoothing using the Matlab-based TIMESAT package [36]. TIMESAT supports three different filters for smoothing time series datasets, i.e., adaptive Savitzky–Golay filter, asymmetric Gaussian filter, and double logistic functions. The GUI controls for this process determine the necessary settings based on the fitness of the smoothed curve to the original time series of vegetation indices. Regarding the choice of filtering method, Jönsson and Eklundh [36], the developers of TIMESAT, and many studies have picked the most appropriate filter in their respective study area depending on the research purpose and the nature of noise in the time series [22,37].

By comparing the fitness of these three filtering methods to the time series vegetation indices, the adaptive Savitzky–Golay filter was finally selected as the filtering algorithm for time-series vegetation indices' smoothing and phenometrics' extraction. The settings of the model parameters determined by the fit effect of the crop sample time series curves are shown in Table 1. As proposed by Richard et al. [38] in extracting the vegetation phenology of the Kenyan region, setting the percentage amplitude parameter of the smoothing function at the season start/end values at 20% can optimize the error caused by varying start and end dates of a season in different locations across the whole study area. Therefore, the same setup was used in our study.

**Table 1.** Optimal parameter setting for curve fitting model.

| Parameters | Settings |
|---|---|
| Curve-fitting model | Savitzky–Golay filtering |
| Seasonality parameter | 0 (0 will attempt to fit two seasons) |
| Spike method | 3 (STL original) |
| No. of envelope iterations | 1 |
| Adaptation strength | 3 |
| Window size | 4 (only for Savitzky–Golay filtering) |
| Season start/end values | 0.2 |

Once the best curve fitting model and its optimal settings were determined, the filtered S-NDVI and S-EVI datasets were then generated from the SITS of vegetation index over the entire study area. From above, as indicators to describe the vegetation phenological phase (i.e., P-NDVI and P-EVI), 13 phenometrics (e.g., the beginning times, the largest value, the length of the seasons, etc.) were extracted from the smoothed vegetation indices profile by TIMESAT software. P-NDVI/EVI 2018 were obtained from MOD13Q1 time series from 2017 to 2019. The meanings of the phenometrics of vegetation index (P-VIs) are as follows:

1. Beginning of the season: The date from the minimum value at the left edge to a user-defined value (usually a proportion of the seasonal amplitude).
2. End of the season: The date from the minimum value at the right edge to the user-defined value.
3. Length of the season: Days from the beginning to the end of the growing season.
4. Base level: The average of the minimum values around the complete growing season.
5. Time for the mid of the season: The average of the dates corresponds to the increase to 80% of the peak and the decrease to 80% of the peak.
6. Largest data value for the fitted function during the season: The peak of the fitted growing season curve.
7. Seasonal amplitude: The difference between the growing season peak and the base value.
8. Left derivative: The ratio of the difference between 20% and 80% of the left peak to the corresponding time difference.
9. Right derivative: The absolute value of the ratio of the difference between 20% and 80% of the peak on the right side and the corresponding time difference.
10. Large seasonal integral: The integral value of the fitted curve from the beginning to the end of the growing season.

11. Small seasonal integral: The integral value of the difference between the fitted curve and the base value from the beginning to the end of the growing season.
12. Value for the beginning of the season: The value of the curve fit corresponding to the beginning of the growing season.
13. Value for the end of the season: The curve-fit value corresponding to the end of the growing season.

#### 2.3.2. Classification Strategy

Appropriate classification strategies should be selected to reduce the dispersion of samples in spectral dimensions and phenological periods, especially when conducting large-scale land use/land cover change studies. In this study, the study area covered entire Kenya and spanned across different climatic zones, resulting in different timings for the onset of the rainy season, especially for the second rainy season (i.e., the short rainy season). Some sample sites across the country typically experience single-peak rainfall (one rainy season), while others experience double-peak rainfall (two rainy seasons) in a calendar year. This variability in the rainy season can lead to differences in phenological parameters among the same vegetation type during the second rainy season [38]. To address this situation, Richard et al. [38] and David et al. [39] proposed using only the phenological parameters extracted during the first growing season in the classification, corresponding to the P-NDVI-1 dataset in this study. However, the above approach disregards the vegetation information of the short rainy season, which could weaken the classification accuracy when identifying crop species. To make full use of all the information provided by the time-series vegetation indices and reduce the heterogeneity of phenological parameters and vegetation spectra among sample sites due to climatic differences, we selected samples within the respective agroecological zones (AEZs) in Kenya only (Figure 1b) for crop classification based on that zone. According to previous studies [18,40,41], these AEZs delineate the production characteristics of the primary sector in different regions of the country and thus contain similar cropping structures and climate conditions within the same zone. These AEZs include four suitable crop areas with significant planting scale: (1) coast, (2) high rainfall, (3) semi-arid north, and (4) semi-arid south; three cropless areas: (1) arid north, (2) arid south, and (3) Turkana; and major metropolitan areas: Nairobi and Mombasa (named cities). This study will verify the rationality of our classification strategy by comparing the accuracy of the independent classification for each AEZ with the national level classification.

#### 2.3.3. Statistical Learning Algorithm

Two widely used advanced statistical learning classifiers, i.e., support vector machine (SVM) and random forest (RF), were selected to investigate the potential of these eight vegetation index time series datasets in crop identification. Briefly, statistical learning refers to the class of algorithms used for classification and regression analysis, including linear and quadratic discriminant analyses, SVM, RF, and neural networks [30]. These classifiers are robust for handling high-dimensional data characterized by strong intercorrelation and high redundancy and have been successfully applied to mapping crops in Africa [42,43]. Previous studies have demonstrated that in crop identification based on time series data, RF is characterized by high accuracy, fast calculation speed, and easy adjustment of parameters compared to other machine learning methods, such as SVM [44,45]. Moreover, RF has been found to produce more robust mapping results in fragmented small-scale farming areas in Africa compared to other methods [46]. However, this result cannot be generalized, e.g., Shao and Lunetta (2012) [47] compared the classification performance of SVM, neural network, and CART algorithms using MODIS time series data and concluded that SVM outperforms other algorithms in land use classification. Therefore, we chose SVM and RF as the classifiers for crop identification in this paper.

The SVM is a classification system derived from statistical learning theory. The SVM algorithm uses non-linear mappings (i.e., kernel function) to project the input vectors to a very-high-dimension feature space, then creates a linear decision surface (i.e., hyperplane, Cortes and Vapnik (1995) [48]) in this new feature space to minimize errors and distinguish data classes. RF is an ensemble learning algorithm based on classification and regression tree (CART), where each of the CART classification trees is trained with some samples and features through random sampling with replacement from the training set and finally generating the final prediction result by the voting principle.

Operationally, the algorithms used in this experiment were based on EnMAP-Box, a QGIS software package developed by the Environmental Mapping and Analysis Program (EMA) project team in Germany. The main hyper-parameters in SVM include kernel function, kernel coefficient $\gamma$, and penalty parameter C. The kernel functions are generally linear, polynomial, radial basis (RBF), and sigmoidal functions. According to the relationship between the number of samples and the number of features in the training set, RBF was selected as the kernel function of the SVM in this study, and the optimal global parameters ($\gamma$ and penalty parameter C) of the classifier were determined by grid search and cross-validation; the parameter setting of RF uses the default parameter in EnMAP-Box (n_estimators = 100).

### 2.3.4. Accuracy Assessment

Five-fold cross-validation was used to evaluate the accuracy of RF classification [49], where the algorithm performs five different training and accuracy evaluations, with 80% of the samples in each training and evaluation being used to train the model, and the remaining 20% of the samples being used to predict the classification accuracy of the model, after which the accuracy of all five classifications is averaged to obtain the 5-fold cross-validation of the final accuracy. As a common method for evaluating the performance of machine learning models, k-fold cross-validation avoids the errors caused by randomness and overfitting of the model by selecting the test set multiple times to evaluate the classification results, thus making the machine learning model more robust.

Overall accuracy (OA) and Kappa coefficient were selected as parameters for accuracy evaluation. Briefly, OA is a fundamental statistic with probabilistic significance for the error matrix. It expresses the probability that the classified result is consistent with the actual type of the area corresponding to the ground for each random sample. The shortfall of OA is that the small change of pixel category may change its percentage, and its objectivity depends on the samples and sampling methods. Kappa analysis uses a discrete multivariate technique, which considers all the factors of the error matrix (misclassification error and leakage error) to overcome the shortcomings of OA. Therefore, OA and Kappa coefficient (Kappa) were generally calculated simultaneously to obtain more accurate information. The OA and Kappa of the classification results for the eight datasets were evaluated using cross-validation to obtain the best preprocessing steps for crop identification and planting area extraction based on MODIS data in Kenya.

### 2.3.5. Variable Importance Analysis

RF classification focusing only on the dominant variables might not significantly reduce accuracy compared to using the full set of SITS features. Performing a variable importance analysis provides some advantages in data storage and computational processing costs and improves the interpretability of the classification model. According to the unique phenological characteristics of crops in Kenya, where the growing season is in sync with the rainy season, an assumption that the SITS of vegetation index during the rainy season has higher importance for the classifier compared to the dry season was introduced in this study. Additionally, the variable importance of the classifier that achieved the best classification results and correlated it with the rainfall data of the corresponding period (at the end of this study) was then calculated to test the assumption. The out-of-bag samples were permuted in the respective variable for each decision tree to calculate the out-of-bag error. The accuracies of the permuted out-of-bag samples were subtracted from the accuracies of the original samples. The average of the differences of the accuracies of a variable gave the raw importance of these variables. Raw variable importance divided by the respective standard deviation gave us the normalized variable importance. It should be noted that the higher the accuracy of the model, the more trustworthy is the importance of the variables. Daily rainfall data for 2018 were obtained from CHIRPS Daily. To maintain consistency with the temporal resolution of the vegetation index time series, the CHIRPS daily rainfall data were aggregated to the same time intervals as the MOD13Q1 MVC to obtain the 16-day cumulative rainfall. Then, we divided them according to AEZ to get the 16-day cumulative average rainfall time series data within each AEZ. A significant linear relationship with the variable importance of the best classification model was found; thus, a linear fit was performed to explore the correlation between them.

## 3. Results

### *3.1. Raw Vegetation Index Time Series, Smoothing Effect, and Phenological Images*

Figure 3a,b shows the R-NDVI time series of Kenya's main land cover types, calculated as the mean of the NDVI values over the randomly selected classification samples. In Figure 3a, the NDVI value of bare land and artificial surface shows no apparent seasonal variation. The curve of bare land was closer to a straight line, with the NDVI value around 0.15. In contrast, the NDVI of the artificial surface was higher than that of bare land, as a mixed image element may contain artificial vegetation such as lawn. The time series of forest & wetland and tea are also relatively straight, but the NDVI value remained at 0.7–0.8, and forest & wetland had a slightly higher NDVI value than tea. The NDVI time series between sisal and wheat was similar. The growth period of sisal is earlier than that of wheat: The NDVI peak of sisal is at the end of April and that of wheat is at the beginning of June. In Figure 3b, sugarcane, maize, rice, herb & shrub have two growing seasons, with the growth period corresponding to the long rainy season from March to May and the short rainy season from October to December, respectively. In addition, maize and herb & shrub have very similar phenological periods, differing only in the overall level of NDVI (the NDVI of maize in all growth seasons is higher than that of herb & shrub). Figure 3c gives the original NDVI time series of double-crop maize and its smoothed effect by adaptive Savitzky–Golay filtering. After smoothing, the irregular sawtooth fluctuation in the curve due to the effects of water vapor, clouds, and aerosols being removed from the curve facilitated the interpretation of phenological periods and computation of phenometrics (Figure 3d).

### *3.2. Classification Accuracy*

Figure A1 presents the learning curve of the RF model at each AEZ when different preprocessing level datasets were input. The learning curve reflected the relationship between out-of-bag accuracy and the number of CART classification trees. At each AEZ, the generalization error decreased as the number of classification trees (n-estimators) increased, and the out-of-bag accuracy increased and converged to the best accuracy that the model could achieve before n-estimators = 100.

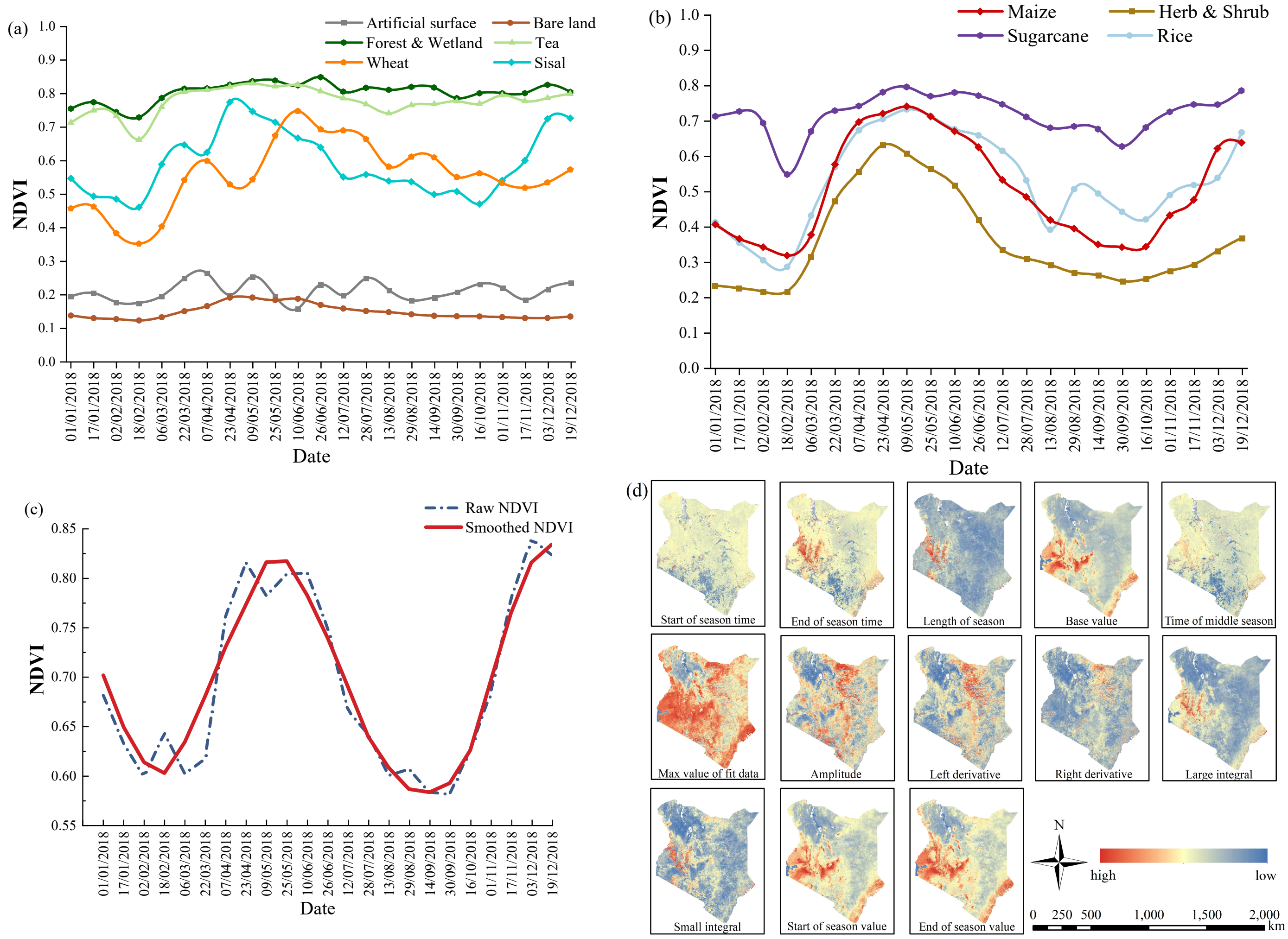

**Figure 3.** (**a**,**b**). Annual time series of the mean MODIS NDVI calculated for major land cover types. (**c**) Example of NDVI time series acquired over a maize field pixel, before and after filtering using Savitzky–Golay algorithm. (**d**) Raster data of phenometrics calculated by TIMESAT (obtained from the first growing season).

Tables 2 and 3 show the OA and Kappa of the RF and SVM classifications obtained from the SITS of the vegetation index dataset with different preprocessing levels at each AEZ. From the box plots of OA and Kappa (Figure 4), both the classification results of RF and SVM showed that S-NDVI-involved classification accuracy was lower than that of R-NDVI, and the classification accuracy of P-NDVI obtained from S-NDVI was the lowest one in the dataset of the three preprocessing levels. The above phenomenon also appeared in the classification results of SITS based on EVI. The addition of the original EVI time series substantially improved the classification accuracy of R-NDVI; however, the addition of the phenological data had minimal effect on the accuracy.

**Table 2.** Accuracies (Overall Accuracy and Kappa Accuracy) were achieved for RF in different AEZs and different datasets.

| | R-NDVI + R-EVI | | R-NDVI + P-NDVI | | R-NDVI | | P-NDVI | | S-NDVI | | R-EVI | | P-EVI | | P-NDVI-1 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** |
| Arid North | 73.78 | 62.27 | 69.85 | 56.88 | 70.7 | 58.03 | 66.35 | 52.22 | 69.72 | 56.65 | 68.09 | 54.13 | 66.64 | 52.6 | 65.48 | 50.88 |
| Arid South | 78.14 | 56.96 | 77.11 | 55.23 | 77.27 | 55.49 | 75.35 | 51.9 | 76.87 | 54.72 | 78.17 | 56.83 | 78.38 | 51.92 | 74.92 | 51.24 |
| Cities | 82.86 | 75.1 | 82.14 | 73.92 | 82.86 | 75.06 | 78.57 | 68.64 | 80.48 | 71.42 | 82.38 | 74.35 | 78.57 | 68.84 | 78.33 | 68.57 |
| Coast | 73.31 | 61.75 | 73.63 | 62.33 | 73.23 | 61.72 | 67.36 | 53.47 | 71.64 | 59.56 | 71.17 | 58.65 | 68.39 | 55 | 63.22 | 47.66 |
| High Rainfall | 73.24 | 57.17 | 68.57 | 49.79 | 68.74 | 49.53 | 65.66 | 46.4 | 67.09 | 48.14 | 70.05 | 51.76 | 65.86 | 46.56 | 63.82 | 44.26 |
| Semi-Arid North | 72.63 | 57.45 | 67.78 | 49.97 | 66.9 | 48.47 | 63.8 | 43.57 | 65.51 | 47.35 | 68.13 | 50.57 | 64.37 | 44.19 | 63.2 | 42.91 |
| Semi-Arid South | 72.99 | 63.17 | 67.5 | 55.76 | 68.02 | 56.19 | 64.79 | 52.35 | 67.01 | 55.38 | 65.02 | 51.9 | 65.23 | 52.86 | 61.62 | 48.22 |
| Turkana | 80.16 | 68.74 | 64.62 | 44.6 | 65.88 | 49.94 | 64.23 | 43.93 | 63.84 | 43.37 | 70.76 | 54.2 | 64.49 | 44.44 | 62.79 | 41.85 |

**Table 3.** Accuracies (Overall Accuracy and Kappa Accuracy) were achieved for SVM in different AEZs and different datasets.

| | R-NDVI + R-EVI | | R-NDVI + P-NDVI | | R-NDVI | | P-NDVI | | S-NDVI | | R-EVI | | P-EVI | | P-NDVI-1 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** | **OA** | **Kappa** |
| Arid North | 74.49 | 63.9 | 66.4 | 52.95 | 67.75 | 54.68 | 64.9 | 50.31 | 69.12 | 56.29 | 64.8 | 50.44 | 64.27 | 50.0 | 64.53 | 49.61 |
| Arid South | 77.41 | 57.62 | 75.85 | 54.5 | 75.58 | 53.93 | 75.65 | 53.16 | 73.59 | 50.97 | 73.95 | 51.49 | 74.62 | 52.0 | 73.36 | 49.22 |
| Cities | 82.62 | 74.72 | 77.86 | 67.77 | 83.1 | 75.38 | 77.38 | 66.91 | 79.52 | 70.12 | 81.9 | 73.81 | 79.29 | 69.64 | 79.05 | 69.4 |
| Coast | 72.99 | 61.69 | 70.85 | 58.72 | 72.76 | 61.02 | 63.86 | 48.51 | 71.8 | 59.9 | 69.9 | 57.03 | 64.65 | 49.66 | 60.92 | 44.39 |
| High Rainfall | 71.76 | 58.4 | 61.59 | 45.52 | 64.79 | 47.3 | 58.89 | 41.44 | 62.55 | 46.11 | 65.37 | 48.95 | 58.82 | 41.71 | 60.37 | 41.06 |
| Semi-Arid North | 73.26 | 59.96 | 68.73 | 52.53 | 66.52 | 49.36 | 63.04 | 44.13 | 67.63 | 50.25 | 67.53 | 50.66 | 62.15 | 42.75 | 61.58 | 41.31 |
| Semi-Arid South | 73.33 | 64.82 | 65.41 | 54.44 | 65.31 | 54.11 | 63.4 | 51.54 | 66.03 | 54.96 | 64.17 | 52.62 | 63.4 | 51.54 | 59.84 | 46.8 |
| Turkana | 78.85 | 66.7 | 65.8 | 46.31 | 67.62 | 49.37 | 63.84 | 43.35 | 65.27 | 45.57 | 71.28 | 54.95 | 63.58 | 42.96 | 61.88 | 40.41 |

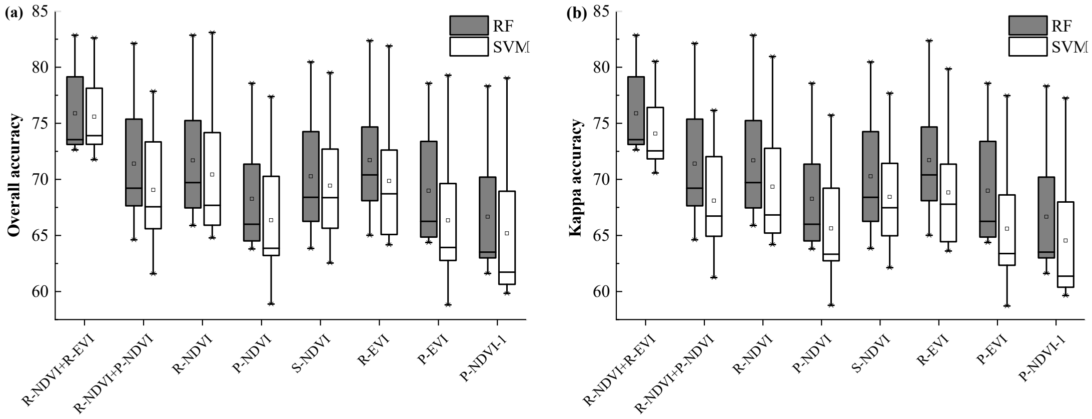

**Figure 4.** Box plot of classification accuracy for each dataset: (**a**) overall accuracy; (**b**) Kappa accuracy.

Due to the difference of SVM and RF classification accuracy on various datasets showing consistency, and because RF classifier had higher performance, we introduced only the RF results. As the most effective dataset for RF classifier, R-NDVI + R-EVI obtained the best results (OA = 82.86%) in cities, followed by Turkana (OA = 80.16%), arid south (OA = 78.14%), arid north (OA = 73.78%), coast (OA = 73.31%), high rainfall (73.24%), and semi-arid south (72.99%); semi-arid north had the lowest classification accuracy, with an OA of 72.63%. The classification maps of each AEZ were mosaicked together to obtain the LULC map of Kenya in 2018 (Figure 5).

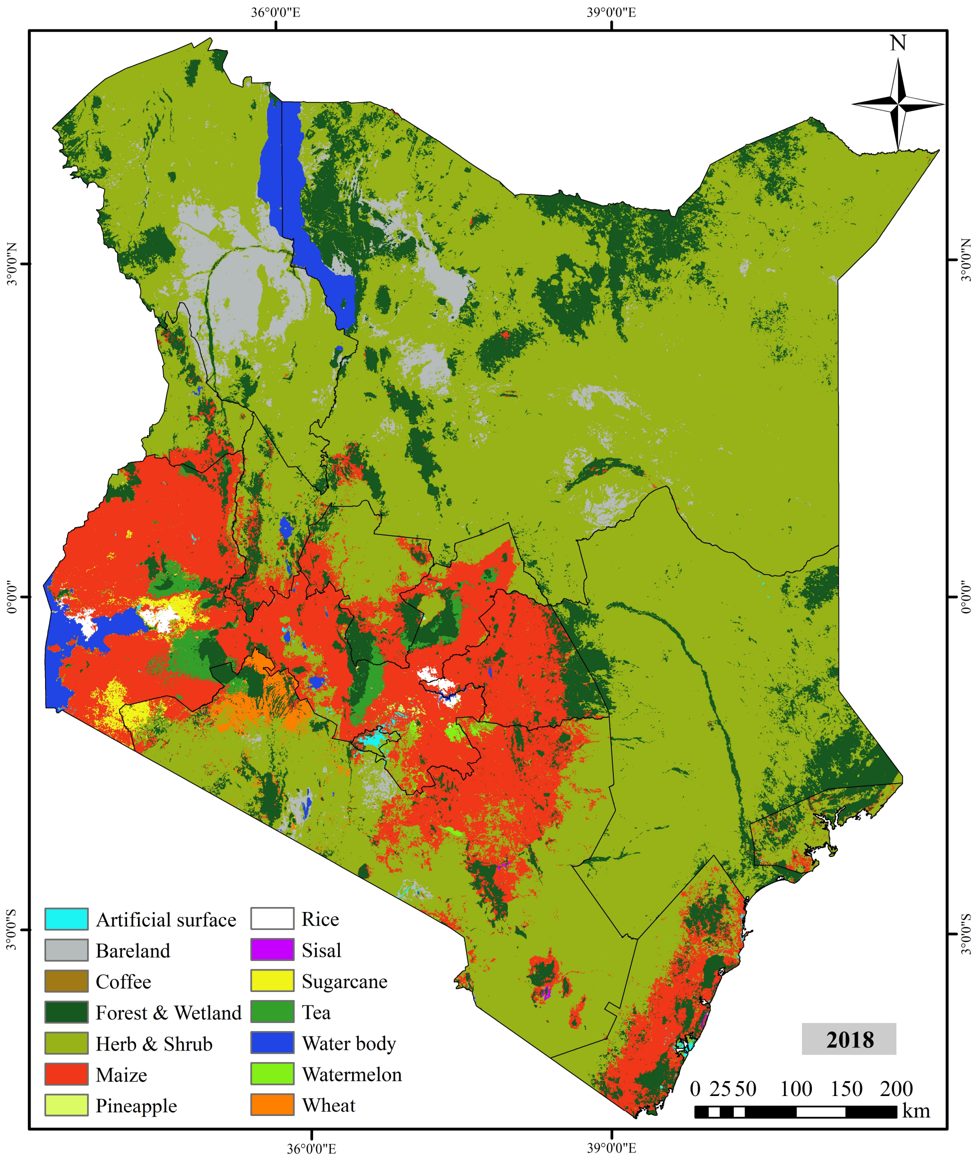

**Figure 5.** RF classification of Kenya in 2018.

### *3.3. Validation of Classification Strategy*

Independent samples within the eight AEZs performed RF classification and 5-fold cross-validation on the vegetation index dataset R-NDVI + R-EVI for entire Kenya. The final classification results (OA = 62.82%, Kappa = 50.52%) were lower than that of any AEZ. Figure 6 illustrates the common differences between the two classification strategies in the entire study area by taking a typical case of extracting tea and wheat in the northeast of Narok County. Figure 6a shows the results of independent classification of each AZE. Figure 6b presents the results of RF classification at the national level. Figure 6c displays the Google Earth high-resolution image of the area, where the distribution of tea and wheat was provided by Crop mask 2015. As can be seen in Figure 6a,c, the tea-growing area (in the black-highlighted part) and the wheat area (in the white-highlighted part) were successfully classified. In contrast, Figure 6b shows that the RF classification misidentified the crop types of the two mentioned growing areas as maize.

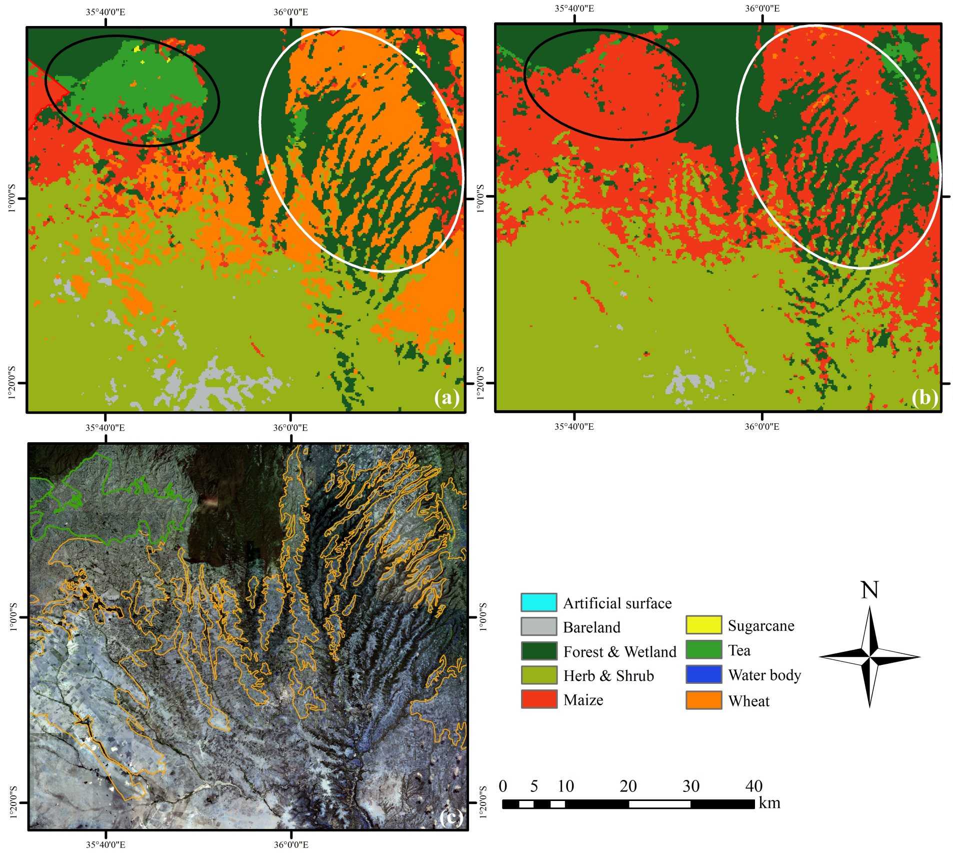

**Figure 6.** Comparison of the two classification strategies in extracting tea and wheat in northeastern Narok County. (**a**) The results of independent classification of each AZE; (**b**) the results of RF classification at a national scale; (**c**) Google Earth high-resolution image of the area. Note that the distribution of tea and wheat was provided by Crop mask 2015.

### 3.4. Planting Area Extraction

Crop acreage estimation was calculated using the classification strategies with the SITS of vegetation indices in 2000 and 2015 in Kenya and compared with the farmland area provided by the Kenya Crop Mask (RCMRD, 2018) (Figure 7). Figure 7 shows that the RF classifier overestimated the acreage of maize and sugarcane and gave various degrees of underestimation of tea, wheat, rice, and watermelon.

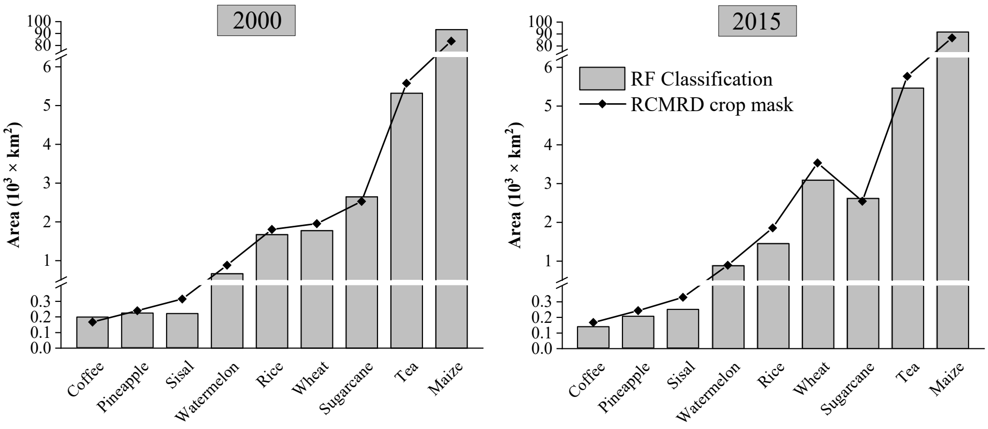

**Figure 7.** MOD13Q1: Acreage estimation of coffee, pineapple, sisal, watermelon, rice, wheat, sugarcane, tea, and maize in Kenya using RF classification and RCMRD crop mask.

The mean absolute percentage error (MAPE) was calculated for the area of each crop in 2000 and 2015:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{actual_t - forecast_t}{actual_t}\right| \times 100\% \tag{1}$$

where *actual* is the crop acreage from Crop Mask, *forecast* is the predicted area obtained from the RF model, *n* is the number of observations, and t is the iteration of observations. MAPE is an average of the absolute percentage errors from model predictions, i.e., the average of the ratio of the absolute area error with the actual area.

The MAPE values of maize, wheat, rice, tea, sugarcane, watermelon, sisal, pineapple, and coffee were 8.5%, 10.7%, 14.2%, 4.9%, 4.0%, 12.1%, 27%, 10.7%, and 17.8%, respectively. Most of the main crops were well identified with MAPE no more than 15%. The classification method proposed in this study is suitable for mapping the distribution of major crops in Kenya. Therefore, we applied the model to extract the distribution of major crops in Kenya in 2020. The area of major crops is shown in Table 4 using pixel statistic models.

**Table 4.** Area distribution of major crops in Kenya in 2020 obtained using RF classifier (unit: km$^2$).

| | **Coffee** | **Maize** | **Rice** | **Sugarcane** | **Tea** | **Watermelon** | **Wheat** | **Sisal** | **Pineapple** |
|---|---|---|---|---|---|---|---|---|---|
| Area (km$^2$) | 169.96 | 93006.65 | 1600.44 | 2493.80 | 5531.54 | 743.04 | 3491.75 | 253.40 | 191.31 |

### 3.5. Variable Importance Analysis

According to the distribution of crops in Kenya (Figure 5), the high-rainfall area, which has the wealthiest crop types, was selected for the variable importance analysis because rich features place a higher demand on the discriminatory ability of the classifier variables. Figure 8 presents the QA summary images of the SITS from MOD13Q1 during the long and short rainy seasons of 2018 in the high-rainfall area. Benefitting from MVC, high-quality images were obtained in this region during both the long and short rainy seasons and enabled us to explore the contribution difference of vegetation index time series in classification between various seasons. The variable importance of NDVI time series in the model as histograms and the superimposed 16-day cumulative average rainfall time series data are shown in Figure 9a. The importance of NDVI variables (see Figure 9a) demonstrated a clear consistency with rainfall in both the long and short rainy seasons. From the onset of the rainy season, crops enter a rapid growth phase. The differences between the time series of vegetation indices of each crop showed explicitly that the importance of vegetation index variables increased significantly in this period compared with the dry season. To quantify the correlation between the season and the categorical contribution of the vegetation index in the corresponding time, a linear fitting between rainfall and the importance of NDVI variables was conducted (see Figure 9b). In the high-rainfall region, rainfall experienced a significant positive linear relationship ($p < 0.05$), with NDVI variable importance of $r^2 = 0.35$.

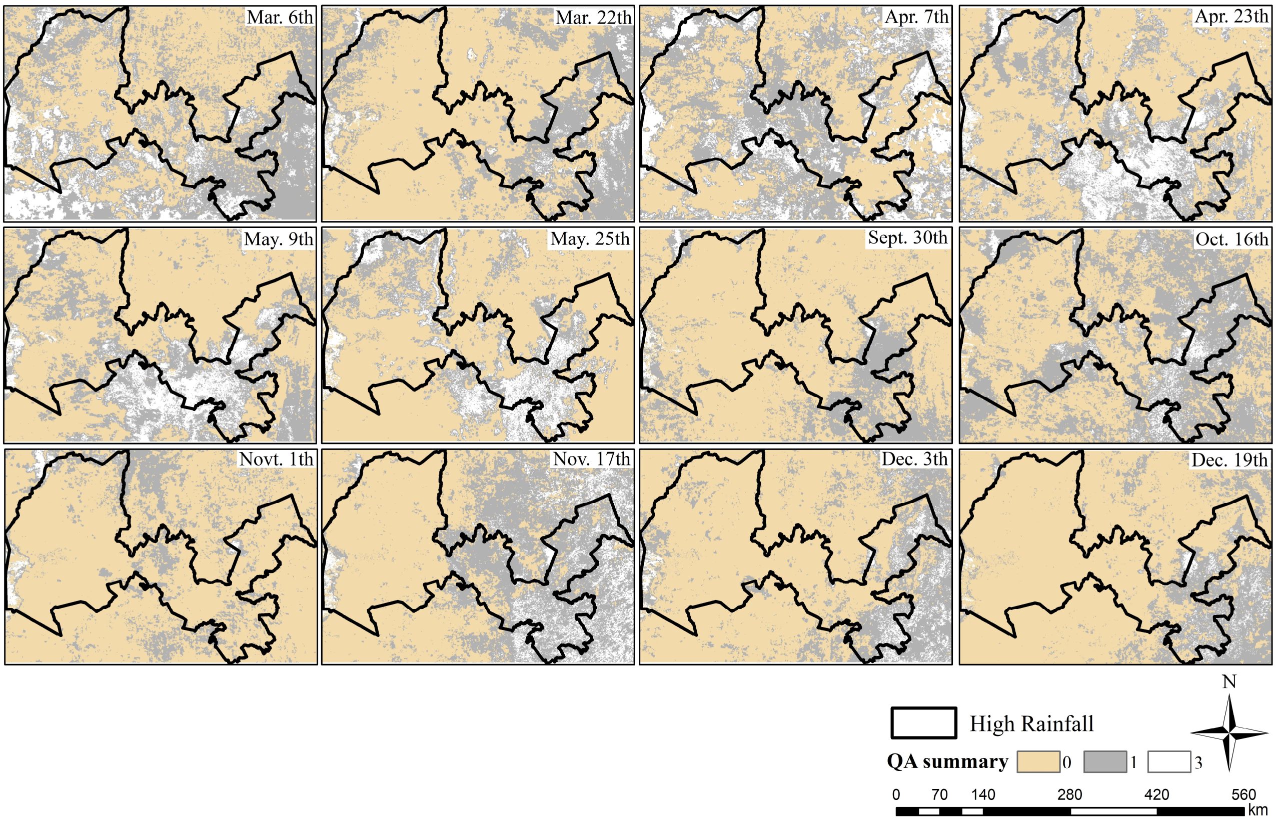

**Figure 8.** MOD13Q1 QA summary for the high-rainfall area in the long rainy season and short rainy season. 0: good data, use with confidence, 1: useful marginal data, 3: pixel fully covered by clouds. Only three images had more than 10% cloud cover during the rainy season: 23 April, 9 May, and 25 May (15.3%, 15.1%, and 13.7%, respectively).

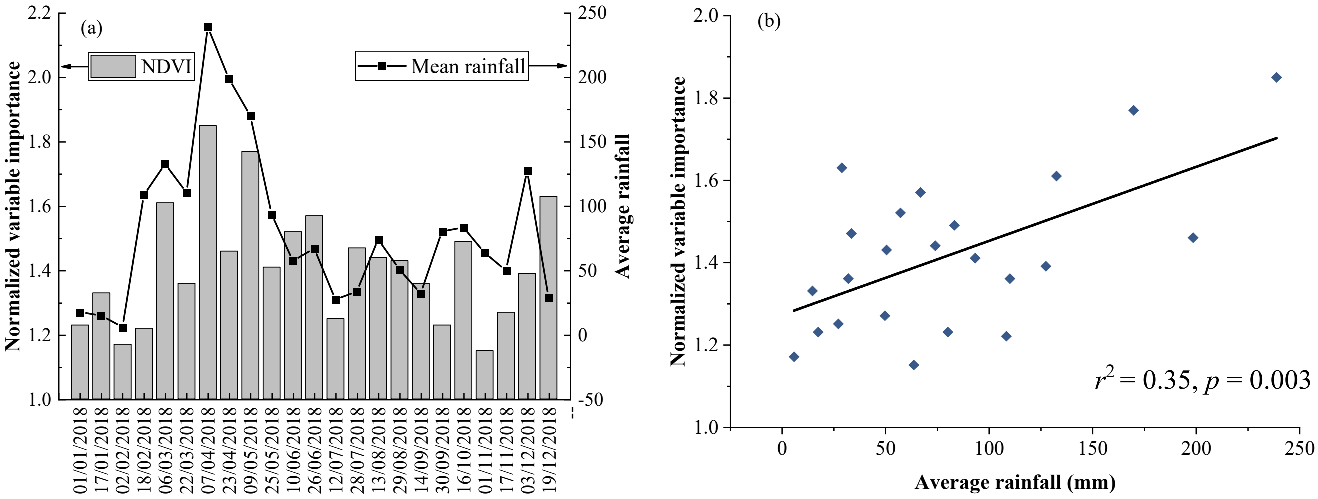

**Figure 9.** (**a**) Normalized variable importance of NDVI and 16-day cumulative mean rainfall data in the high-rainfall area and (**b**) the linear relationship between them.

## 4. Discussion

Among the eight VI datasets, the highest mean (RF OA = 75.89%, RF Kappa = 62.82%) and median (RF OA = 73.55%, RF Kappa = 60.01%) classification accuracies were obtained for the R-NDVI + R-EVI involved in each AEZ. The original vegetation index time series (R-NDVI, R-EVI) outperformed the smoothed time series (S-NDVI) to some extent and also performed better than the classification accuracy of the datasets of phenological parameters

(P-NDVI, P-EVI) obtained from the respective vegetation indices. Accuracy assessments showed that the adaptive Savitzky–Golay filter reduced the mean OA and Kappa of the original NDVI time series by 1.43% and 2.23%, respectively, in eight AEZs. This outcome demonstrated that original vegetation indices can provide more information than smoothed vegetation indices for advanced statistical learning classifiers in the Kenyan region, which is in agreement with the results of Araujo Picoli et al. [30]. Our result is also consistent with the findings of Kuchler et al. [29] in mapping the main cropping systems in the northern region of the state of Togrosso, Amazonas, Brazil, where the authors achieved the highest classification accuracy in both RF and SVM classifications using original MODIS vegetation index products without noise reduction processing. Chen et al. [27] suggested that Savitzky–Golay filtering reduces the interference from clouds and atmosphere in the vegetation index time series on the one hand but also filters out important information, on the other hand, thus hindering the separation of crop types because of their spectral similarity. The drawback mentioned above is particularly significant when using an RF classifier, which has the advantage of identifying the most discriminative variables when dealing with strongly intercorrelated data. When smoothing the time series using a filtering algorithm, discriminative information from one band (temporal phase) of the time series can be transferred to adjacent bands. Moreover, in our case, two rainy seasons per year imply high cloud coverage, and the critical discriminative information is likely to be corrupted in the process of removing the cloud interference.

In addition, despite reducing the data input by half, the classification accuracy of P-NDVI-1 was not significantly lower than that of two full growing seasons' phenology (P-NDVI), proving that the proposal of Richard et al. [38] and David et al. [39] using only the first growing season's phenology parameters for crop classification is reasonable. Nevertheless, neither of these types of phenological information can meet the requirements because the dataset always performs worse in the classification (RF accuracy of P-NDVI: mean OA = 68.26%, mean Kappa = 51.56%; RF accuracy of P-NDVI-1: mean OA = 66.67%, mean Kappa = 49.45%). The reason for the low classification accuracy of the phenological dataset might be the similar VI profiles of various crops in Kenya, where most vegetation types (including crops, grassland and shrubs, and forest) enter the growing season rapidly after the onset of the rainy season, and the phenological differences among crops are not sufficient for crop discrimination, i.e., the phenological data from the vegetation index reflect seasonal changes more than differences in physiological processes for vegetation.

Furthermore, a comparison of the classification accuracies of R-NDVI and R-EVI found similar identification abilities of these vegetation indices in the study area. In the Turkana area, R-EVI slightly outperformed R-NDVI (RF OA = 64.88%, RF Kappa = 49.94%) with RF OA = 70.76% and RF Kappa = 54.2%, probably due to the canopy background adjustment factor L in EVI, which weakened the sensitivity of EVI to most canopy backgrounds except snow [50]. Thus, R-EVI achieved higher classification accuracy in Turkana, where sparse grassland and bare ground are dominant types. In summary, the conclusion reached by Hastie [51] was confirmed in a study of crop classification in the Kenyan region: The machine learning algorithm based on statistical learning theory is more likely to yield the best results when using raw datasets than smoothed data.

Although Semi-Arid North and Semi-Arid South have relatively simple land cover types in eight AEZs (mainly herb & shrub, maize, and forest), their OA in the dataset that achieved the highest accuracy (i.e., R-NDVI + R-EVI) was less than 75%. This result was expected, as shown in Figure 3b, where maize and herb & shrub have very similar NDVI time series curves, showing only numerical differences. We found that these two AEZs have lower planting densities than other regions. Specifically, according to the maize yield data published by the Kenya Ministry of Agriculture, Livestock, Fisheries and Irrigation (MOALFI), harvested areas in Makueni and Kitui (major maize planting counties of Semi-Arid North and Semi-Arid South) were as high as 137,330 ha and 83,177 ha, respectively. In contrast, yields were only 0.56 mt/ha and 0.36 mt/ha, respectively. Large regions of sparsely planted maize growing areas are reflected as low values on NDVI, which reduces the difference between maize and herb & shrubs in the NDVI time series, leading to their misclassification by RF.

The misclassification of crop types into maize (refer to Figure 6) is not unexpected in the results of training and classification at the national scale. The phenomenon above may be attributable to several factors. The first one is related to spectral heterogeneity among the same vegetation types due to differences in land conditions, climatic conditions, and cropping habits. With maize being the most widely grown crop in Kenya, the above differences can make it difficult for its vegetation index time series to stay consistent across the country. Additionally, statistical learning classifiers trained using highly heterogeneous samples showed low sensitivity in recognition of other crops during the classification process, manifested by misclassifying a large number of image elements as maize. The second one is related to the class imbalance problem in machine learning. To ensure the training effect, the classification samples should be evenly distributed throughout the study area. Due to the disparity in the scale of cultivation between different crops, the number of maize samples across Kenya differs by up to 500 times from that of other crops (e.g., coffee), and the classifier based on the national scale inevitably faces the sample imbalance problem, which mainly exists in supervised machine learning. The statistical model, with overall classification accuracy as the learning goal, focused too much on the majority class, thus degrading the classification performance of the minority class samples. Through cross-validation and comparison of the classification effects, we demonstrated that the influence of spectral heterogeneity among the same vegetation types due to differences in land conditions, climate conditions, and cropping habits on the classification results could be solved to a certain extent through a simple classification strategy. The AEZ division balances the ratio of samples in the training set, effectively solving the challenge of training RF classifiers under class imbalance.

It can be deduced that the errors in the extraction of cropland area using the RF classifier, e.g., the overestimation of maize area, might originate from the mixed pixel of MODIS images. The spatial resolution of MODIS imagery and the agricultural scale of smallholder farms in Kenya lead to a high degree of heterogeneity of features exhibited within a single pixel, i.e., the phenomena of mixed pixels are particularly severe. Furthermore, mixed cropping and intercropping could also increase the spectral heterogeneity of mixed pixels, e.g., maize, which is widely grown in Kenya, is usually mixed with beans, cowpeas, green grams, sorghum, millet, and wheat. The spectral characteristics of mixed and intercropping areas are more complex than those of mono-cropping areas, making it easier to misclassify the image pixels in the area. However, the selection of MODIS imagery was dictated by both the research objectives and the requirements of image spatio-temporal resolution in the Kenyan region. On the one hand, as mentioned in the introduction, data source selection can be limited by cloudiness, the rapid response of crops to the rainy season, and the scale of the study area. Through the MVC, the maximum value of vegetation indices was extracted for each pixel over every 16-day period in MODIS daily observation to avoid cloud pollution. Traditional high-spatial-resolution images would fail to meet the demand in terms of temporal resolution (e.g., the revisit period of Sentinel-2 is 5 days with two satellites), and commercial satellites with high spatial and temporal resolution require high input costs. The limitation of swath width also made a vast number of images required for crop monitoring of the above satellites at a national scale, which significantly increased the processing costs. On the other hand, Kenya urgently needs timely crop distribution mapping. Compared with the improvement of the classification accuracy of particular farmland, policy makers might pay more attention to the dynamic changes of crop distribution in the country. We believe the approach proposed in this study can detect nationwide LULC change at a low cost, which is meaningful for countries lacking financial support, such as Kenya.

The variable importance analysis confirmed the conjecture that in Kenya, where the crop growing season overlaps with the rainy season, the vegetation index time series of the rainy season is more important (compared to the dry season) for crop classification using machine-learning algorithms based on statistical learning theory. Besides, it also provides us with a direction for future research by posing the question of whether a machine-learning algorithm can be achieved using only the vegetation index time series during the rainy season for the purpose of reducing the computational cost.

It is also noteworthy that the original data selected in this study is a MODIS vegetation index 16-day synthetic product, which has been widely used for crop identification, growth monitoring, and planting area extraction around the world. However, in a study on temporal sampling for monitoring vegetation condition, Alexandridis et al. [52] indicated that the optimal time resolution for managing vegetation is 7 days, a time interval close to that of the MODIS surface reflectance 8-day synthetic product. In future studies, it would be interesting to study the usage of the MOD09Q1 surface reflectance product to improve the temporal resolution of the NDVI time series from 16 days to 8 days and explore the ability of more frequent observations to detect differences in spectral profiles with similar phenological crops. When estimating crop acreage at the national scale, the application of the dataset that achieved the best performance (i.e., R-NDVI + R-EVI) could also result in relatively poor classification accuracy in some parts of Kenya (e.g., OA = 72.63% in the semi-arid north region). One of the critical reasons for that is related to the cropping pattern in Kenya. The dominant cropping pattern in this region is smallholder farming with a fragmented distribution of cropland, where the landscape contains a high degree of heterogeneity. Mixed pixels containing multiple land cover types in a single pixel usually dominate satellite images [5]. Therefore, more research is needed to investigate the mixed pixel problem in crop classification and planting pattern recognition in large-scale areas with arable patch sizes smaller than the pixel size of MODIS data.

## 5. Conclusions

In this study, the effectiveness of different preprocessing methods of the MODIS time series for mapping the distribution of major crops in Kenya was investigated, and the most appropriate crop distribution extraction algorithm for the region was obtained. The reliability of the method was then tested at the statistical area level. Results demonstrated that, compared to the application of the original vegetation index time series, smoothed vegetation index time series and vegetation phenometrics reduce the classification accuracy of RF and SVM. Modern statistical learning classifiers (such as RF and SVM) exhibit excellent robustness when handling high-dimensional data characterized by strong intercorrelation and high redundancy and can produce satisfactory results when using original datasets containing noise in the Kenyan region. Finally, the relationship between the importance of variables in the best RF classifier and the unique climatic characteristics of Kenya was analyzed, and the results of the linear fit showed a significantly positive correlation between rainfall and the NDVI importance of variables, demonstrating the high significance of the time series of the vegetation index during the rainy season for the RF classifier in Kenya, where the crop-growing season and the flush period overlap. Considering the reduction in labor and computational cost, we believe that the accuracy of the model is reasonable for the extraction of major crops' distribution in the Kenyan region. Therefore, we propose the application of the time-series dataset of the original NDVI overlaid with EVI for crop identification using the statistical learning classifier in the Kenya region. For large-scale regional crop identification, we recommend dividing the monitoring area with available auxiliary data (such as climate information and agricultural zoning) before performing the classification, and each zone should be classified independently.

This study can be used as a basis for accurate crop identification in Kenya. All the sample data used in this study were derived from high-resolution Google Earth imagery paired with Crop mask files. The classification results can be further improved by considering additional ground truthing data.

**Author Contributions:** Conceptualization, R.N. and C.H.; methodology, R.N.; validation, R.N., C.H. and Y.L.; resources, Y.L. and X.L.; data curation, R.N., Y.L. and X.L.; writing—original draft preparation, R.N.; writing—review and editing, R.N., X.Z., Y.L., D.M.M., X.L., T.C., W.D. and C.Z.;

project administration, C.H.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by International Partnership Program of Chinese Academy of Sciences, grant number 151542KYSB20200017 and "Strategic Priority Research Program" of CAS, grant number XDA24040312.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets presented in this study are publicly available. MOD13Q1 and CHIRPS Daily are openly available via the Google Earth Engine. The Kenya Crop Mask 2000 and 2015 were produced by Landsat 8 and distributed by the Region Centre for Mapping of Resources for Development (RCMRD) via https://opendata.rcmrd.org/datasets/kenya-crop-mask-2015/explore (accessed on 29 June 2021).

**Acknowledgments:** The authors would like to thank all their colleagues for the fruitful discussions on this work and wish to express their gratitude to USGS and GEE platform for supplying MODIS and CHIRPS data. The authors also sincerely thank the anonymous reviewers for their very competent comments and helpful suggestions. The authors would like to express their gratitude to EditSprings (https://www.editsprings.com/, accessed on 29 June 2021) for the expert linguistic services provided.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

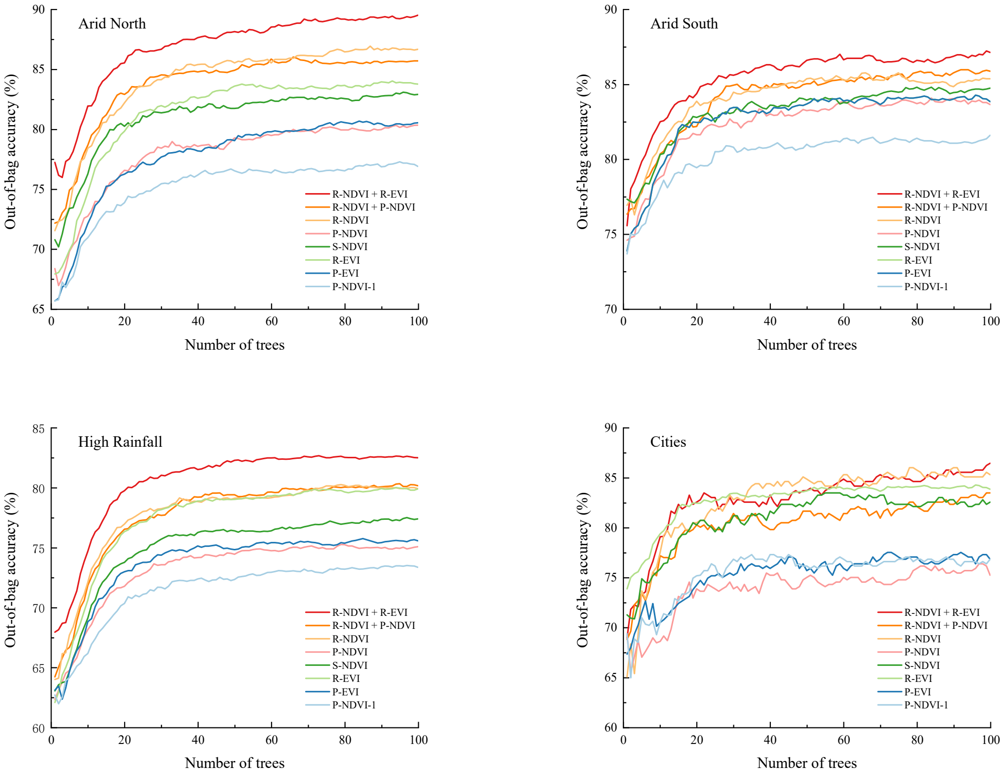

**Figure A1.** *Cont.*

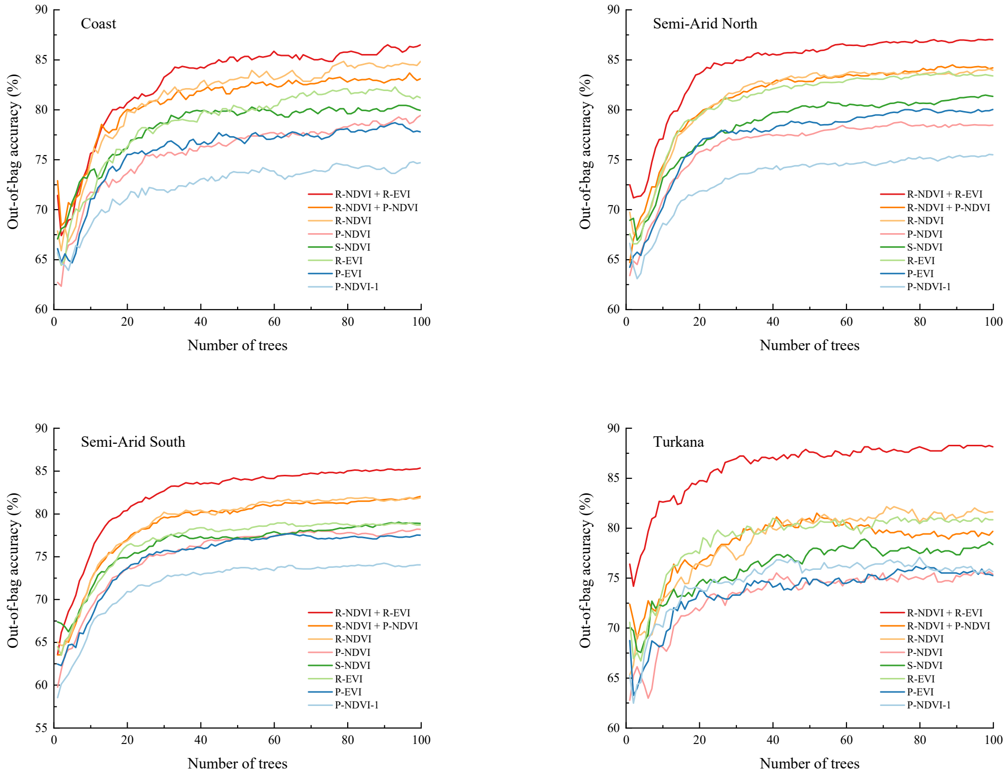

**Figure A1.** Learning curves of RF classification for each AEZ. With an increasing number of trees, the out-of-bag accuracy increases and converges to a threshold.

## References

1. Luciani, R.; Laneve, G.; Jahjah, M. Agricultural Monitoring, an Automatic Procedure for Crop Mapping and Yield Estimation: The Great Rift Valley of Kenya Case. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2196–2208. [CrossRef]
2. FAO. *Guidelines on the Use of Remote Sensing Products to Improve Agricultural Crop Production Forecast Statistics in Sub-Saharan African Countries*; FAO: Rome, Italy, 2018. [CrossRef]
3. Lowder, S.K.; Skoet, J.; Raney, T. The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide. *World Dev.* **2016**, *87*, 16–29. [CrossRef]
4. Samberg, L.H.; Gerber, J.; Ramankutty, N.; Herrero, M.; West, P. Subnational distribution of average farm size and smallholder contributions to global food production. *Environ. Res. Lett.* **2016**, *11*, 124010. [CrossRef]
5. Smith, J.H.; Stehman, S.V.; Wickham, J.D.; Yang, L. Effects of landscape characteristics on land-cover class accuracy. *Remote Sens. Environ.* **2003**, *84*, 342–349. [CrossRef]
6. Piiroinen, R.; Heiskanen, J.; Mõttus, M.; Pellikka, P. Classification of crops across heterogeneous agricultural landscape in Kenya using AisaEAGLE imaging spectroscopy data. *Int. J. Appl. Earth Obs. Geoinform.* **2015**, *39*, 1–8. [CrossRef]
7. Jin, Z.; Azzari, G.; Burke, M.; Aston, S.; Lobell, D.B. Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. *Remote Sens.* **2017**, *9*, 931. [CrossRef]
8. Mosomtai, G.; Odindi, J.; Abdel-Rahman, E.M.; Babin, R.; Fabrice, P.; Mutanga, O.; Tonnang, H.E.Z.; David, G.; Landmann, T. Landscape fragmentation in coffee agroecological subzones in central Kenya: A multiscale remote sensing approach. *J. Appl. Remote Sens.* **2020**, *14*, 044513. [CrossRef]
9. Richard, K.; Abdel-Rahman, E.M.; Subramanian, S.; Nyasani, J.O.; Thiel, M.; Jozani, H.; Borgemeister, C.; Landmann, T. Maize Cropping Systems Mapping Using RapidEye Observations in Agro-Ecological Landscapes in Kenya. *Sensors* **2017**, *17*, 2537. [CrossRef]
10. Maingi, J.K.; Marsh, S.E. Assessment of environmental impacts of river basin development on the riverine forests of eastern Kenya using multi-temporal satellite data. *Int. J. Remote Sens.* **2001**, *22*, 2701–2729. [CrossRef]
11. Tottrup, C. Sensing, Improving tropical forest mapping using multi-date Landsat TM data and pre-classification image smoothing. *Int. J. Remote Sens.* **2004**, *25*, 717–730. [CrossRef]
12. Hashim, M.; Pour, A.B.; Onn, C.H. Optimizing cloud removal from satellite remotely sensed data for monitoring vegetation dynamics in humid tropical climate. *IOP Conf. Ser. Earth Environ. Sci.* **2014**, *18*, 12010. [CrossRef]
13. Vithanage, J.; Miller, S.N.; Driese, K. Land cover characterization for a watershed in Kenya using MODIS data and Fourier algorithms. *J. Appl. Remote Sens.* **2016**, *10*, 045015. [CrossRef]

14. Baldyga, T.J.; Miller, S.N.; Driese, K.L.; Gichaba, C.M. Assessing land cover change in Kenya's Mau Forest region using remotely sensed data. *Afric. J. Ecol.* **2008**, *46*, 46–54. [CrossRef]
15. Moody, A.; Johnson, D.M. Land-Surface Phenologies from AVHRR Using the Discrete Fourier Transform. *Remote Sens. Environ.* **2001**, *75*, 305–323. [CrossRef]
16. Mwaniki, W.M.; Möller, S.M. Knowledge based multi-source, time series classification: A case study of central region of Kenya. *Appl. Geogr.* **2015**, *60*, 58–68. [CrossRef]
17. Luciani, R.; Laneve, G.; Jahjah, M. Developing a classification method for periodically updating agricultural maps in Kenya. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*; IEEE: Beijing, China, 2016; pp. 3543–3546. [CrossRef]
18. Luciani, R.; Laneve, G.; Jahjah, M.; Collins, M. Crop species classification: A phenology based approach. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*; IEEE: Fort Worth, TX, USA, 2017; pp. 4390–4393. [CrossRef]
19. Gachoki, S.M. Estimating Vegetation Phenology at 30m Resolution with Multi-Temporal Optical Imagery for a Rangeland Site in Kenya. Master's Thesis, University of Twente, Enschede, The Netherlands, 2018.
20. Jin, Z.; Azzari, G.; You, C.; Di Tommaso, S.; Aston, S.; Burke, M.; Lobell, D.B. Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sens. Environ.* **2019**, *228*, 115–128. [CrossRef]
21. Sun, R.; Chen, S.; Su, H.; Mi, C.; Jin, N. The Effect of NDVI Time Series Density Derived from Spatiotemporal Fusion of Multisource Remote Sensing Data on Crop Classification Accuracy. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 502. [CrossRef]
22. Shao, Y.; Lunetta, R.S.; Wheeler, B.; Iiames, J.S.; Campbell, J.B. An evaluation of time-series smoothing algorithms for land-cover classifications using MODIS-NDVI multi-temporal data. *Remote Sens. Environ.* **2016**, *174*, 258–265. [CrossRef]
23. Brown, J.; Kastens, J.H.; Coutinho, A.C.; Victoria, D.D.C.; Bishop, C.R. Classifying multiyear agricultural land use data from Mato Grosso using time-series MODIS vegetation index data. *Remote Sens. Environ.* **2013**, *130*, 39–50. [CrossRef]
24. Atkinson, P.; Jeganathan, C.; Dash, J.; Atzberger, C. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. *Remote Sens. Environ.* **2012**, *123*, 400–417. [CrossRef]
25. Zhong, L.; Hu, L.; Yu, L.; Gong, P.; Biging, G.S. Automated mapping of soybean and corn using phenology. *ISPRS J. Photogramm. Remote Sens.* **2016**, *119*, 151–164. [CrossRef]
26. Valero, S.; Morin, D.; Inglada, J.; Sepulcre, G.; Arias, M.; Hagolle, O.; Dedieu, G.; Bontemps, S.; Defourny, P.; Koetz, B. Production of a Dynamic Cropland Mask by Processing Remote Sensing Image Series at High Temporal and Spatial Resolutions. *Remote Sens.* **2016**, *8*, 55. [CrossRef]
27. Chen, Y.; Lu, D.; Moran, E.; Batistella, M.; Dutra, L.V.; Sanches, I.D.A.; da Silva, R.F.B.; Huang, J.; Luiz, A.J.B.; de Oliveira, M.A.F. Mapping croplands, cropping patterns, and crop types using MODIS time-series data. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *69*, 133–147. [CrossRef]
28. Pelletier, C.; Valero, S.; Inglada, J.; Champion, N.; Dedieu, G. Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. *Remote Sens. Environ.* **2016**, *187*, 156–168. [CrossRef]
29. Kuchler, P.C.; Bégué, A.; Simões, M.; Gaetano, R.; Arvor, D.; Ferraz, R.P. Assessing the optimal preprocessing steps of MODIS time series to map cropping systems in Mato Grosso, Brazil. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *92*, 102150. [CrossRef]
30. Picoli, M.C.A.; Camara, G.; Sanches, I.; Simões, R.; Carvalho, A.; Maciel, A.; Coutinho, A.; Esquerdo, J.; Antunes, J.; Begotti, R.; et al. Big earth observation time series analysis for monitoring Brazilian agriculture. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 328–339. [CrossRef]
31. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; Volume 103.
32. Muthoni, F.K.; Odongo, V.O.; Ochieng, J.; Mugalavai, E.M.; Mourice, S.K.; Hoesche-Zeledon, I.; Mwila, M.; Bekunda, M. Long-term spatial-temporal trends and variability of rainfall over Eastern and Southern Africa. *Theor. Appl. Climatol.* **2018**, *137*, 1869–1882. [CrossRef]
33. Li, X.; Chen, S.; Aluoch, S.O.; Mosongo, P.S.; Cao, J.; Hu, C. Maize production status and yield limiting factors of Kenya. *Chin. J. Eco-Agric.* **2018**, *26*, 567–573.
34. Zhang, Y.; Song, C.; Band, L.E.; Sun, G.; Li, J. Reanalysis of global terrestrial vegetation trends from MODIS products: Browning or greening? *Remote Sens. Environ.* **2017**, *191*, 145–155. [CrossRef]
35. Funk, C.; Peterson, P.; Landsfeld, M.; Pedreros, D.; Verdin, J.; Shukla, S.; Husak, G.; Rowland, J.; Harrison, L.; Hoell, A.; et al. The climate hazards infrared precipitation with stations—A new environmental record for monitoring extremes. *Sci. Data* **2015**, *2*, 150066. [CrossRef]
36. Jönsson, P.; Eklundh, L. TIMESAT—A program for analyzing time-series of satellite sensor data. *Comput. Geosci.* **2004**, *30*, 833–845. [CrossRef]
37. De Castro, A.I.; Six, J.; Plant, R.E.; Peña, J.M. Mapping Crop Calendar Events and Phenology-Related Metrics at the Parcel Level by Object-Based Image Analysis (OBIA) of MODIS-NDVI Time-Series: A Case Study in Central California. *Remote Sens.* **2018**, *10*, 1745. [CrossRef]
38. Richard, K.; Abdel-Rahman, E.M.; Mohamed, S.A.; Ekesi, S.; Borgemeister, C.; Landmann, T. Importance of Remotely-Sensed Vegetation Variables for Predicting the Spatial Distribution of African Citrus Triozid (Trioza erytreae) in Kenya. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 429. [CrossRef]

39. Makori, D.M.; Fombong, A.T.; Abdel-Rahman, E.M.; Nkoba, K.; Ongus, J.; Irungu, J.; Mosomtai, G.; Makau, S.; Mutanga, O.; Odindi, J.; et al. Predicting Spatial Distribution of Key Honeybee Pests in Kenya Using Remotely Sensed and Bioclimatic Variables: Key Honeybee Pests Distribution Models. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 66. [CrossRef]
40. Vigani, M.; Dudu, H.; Solano-Hermosilla, G. *Estimation of Food Demand Parameters in Ethiopia: A Quadratic Almost Ideal Demand System (QUAIDS) Approach*; Joint Research Centre: Ispra, Italy, 2019.
41. Mabiso, A.; Pauw, K.; Benin, S. Agricultural growth and poverty reduction in Kenya: Technical analysis for the Agricultural Sectoral Development Strategy (ASDS)—Medium Term Investment Plan (MTIP). *Reg. Strateg. Anal. Knowl. Support Syst. (ReSAKSS) Work. Pap.* **2015**, *35*. Available online: https://ebrary.ifpri.org/utils/getfile/collection/p15738coll2/id/127063/filename/127274.pdf (accessed on 10 December 2021).
42. Forkuor, G.; Conrad, C.; Thiel, M.; Landmann, T.; Barry, B. Evaluating the sequential masking classification approach for improving crop discrimination in the Sudanian Savanna of West Africa. *Comput. Electron. Agric.* **2015**, *118*, 380–389. [CrossRef]
43. Forkuor, G.; Conrad, C.; Thiel, M.; Ullmann, T.; Zoungrana, E. Integration of Optical and Synthetic Aperture Radar Imagery for Improving Crop Mapping in Northwestern Benin, West Africa. *Remote Sens.* **2014**, *6*, 6472–6499. [CrossRef]
44. Inglada, J.; Arias, M.; Tardy, B.; Hagolle, O.; Valero, S.; Morin, D.; Dedieu, G.; Sepulcre, G.; Bontemps, S.; Defourny, P.; et al. Assessment of an Operational System for Crop Type Map Production Using High Temporal and Spatial Resolution Satellite Optical Imagery. *Remote Sens.* **2015**, *7*, 12356–12379. [CrossRef]
45. Onojeghuo, A.O.; Blackburn, G.A.; Wang, Q.; Atkinson, P.M.; Kindred, D.; Miao, Y. Mapping paddy rice fields by applying machine learning algorithms to multi-temporal Sentinel-1A and Landsat data. *Int. J. Remote Sens.* **2017**, *39*, 1042–1067. [CrossRef]
46. Mudereri, B.T.; Dube, T.; Niassy, S.; Kimathi, E.; Landmann, T.; Khan, Z.; Abdel-Rahman, E.M. Is it possible to discern Striga weed (Striga hermonthica) infestation levels in maize agro-ecological systems using in-situ spectroscopy? *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *85*, 102008. [CrossRef]
47. Shao, Y.; Lunetta, R.S. Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. *ISPRS J. Photogramm. Remote Sens.* **2012**, *70*, 78–87. [CrossRef]
48. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
49. Wiens, T.S.; Dale, B.C.; Boyce, M.S.; Kershaw, G.P. Three way k-fold cross-validation of resource selection functions. *Ecol. Model.* **2008**, *212*, 244–255. [CrossRef]
50. Gillieson, D.; Lawson, T.; Searle, L. Applications of High Resolution Remote Sensing in Rainforest Ecology and Management. *Living A Dyn. Trop. For. Landsc.* **2008**, 334–348. [CrossRef]
51. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media: New York, NY, USA, 2009.
52. Alexandridis, T.K.; Gitas, I.; Silleos, N.G. An estimation of the optimum temporal resolution for monitoring vegetation condition on a nationwide scale using MODIS/Terra data. *Int. J. Remote Sens.* **2008**, *29*, 3589–3607. [CrossRef]