remote sensing

MDPI

Article

Sentinel-2 Images Based Modeling of Grassland Above-Ground Biomass Using Random Forest Algorithm: A Case Study on the Tibetan Plateau

Xinyue Fan [1,2], Guojin He [1,2,3,*], Wenyi Zhang [1], Tengfei Long [1,2], Xiaomei Zhang [1], Guizhou Wang [1], Geng Sun [4], Huakun Zhou [5], Zhanhuan Shang [6], Dashuan Tian [7], Xiangyi Li [2,8] and Xiaoning Song [2]

1 Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
2 University of Chinese Academy of Sciences, Beijing 100049, China
3 Key Laboratory of Earth Observation of Hainan Province, Hainan Research Institute, Aerospace Information Research Institute, Chinese Academy of Sciences, Sanya 572029, China
4 Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu 610041, China
5 Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810008, China
6 College of Ecology, Lanzhou University, Lanzhou 730000, China
7 Institute of Geographic Csiences and Natureal Resources Research, Chinese Academy of Sciences, Beijing 100101, China
8 Cele National Station of Observation and Research for Desert Grassland Ecosystems, Cele 848300, China
* Correspondence: hegj@aircas.ac.cn; Tel.: +86-010-8217-8190

Abstract: Accurate information on grassland above-ground biomass (AGB) is critical to better understanding the carbon cycle and conserve grassland resources. As a climate-sensitive key ecological function area, it is important to accurately estimate the grassland AGB of the Tibetan Plateau. Sentinel-2 (S2) images have advantages in reducing mixed pixels and the scale effect for remote sensing, while the data volume is correspondingly larger. In order to improve the estimation accuracy while reducing the data volume required for AGB estimation and improving the computational efficiency, this study used the Recursive Feature Elimination (RFE) algorithm to find the optimal feature set and compared the performance of the Cubist, Gradient Boosting Regression Tree (GBRT), random forest (RF) and eXtreme Gradient Boosting (XGBoost) algorithms for estimating AGB. In this study, ten S2 bands, ten S2-derived vegetation indexes, 218 pieces of AGB field survey data, four types of meteorological data and three types of topographic data were used as the alternative input features for the AGB estimation model. The impurity and permutation importance were used as the feature importance calculation method input to the RFE, and the Cubist, GBRT, RF and XGBoost algorithms were used to construct the AGB estimation models. The results showed that the RF algorithm based on the monthly average temperature (T), elevation, Normalized Difference Phenology Index (NDPI), Normalized Difference Infrared Index (NDII) and Palmer Drought Severity Index (PDSI) performed best (R^2 = 0.8838, RMSE = 35.05 g/m^2, LCCC = 2.44, RPPD = 0.91). The above findings suggest that the RF model based on the features related to temperature, altitude, humidity and leaf water content is beneficial to estimate the grassland AGB on the Tibetan Plateau.

Keywords: above-ground biomass; Tibetan Plateau; random forest; recursive feature elimination; alpine grassland

check for updates

Citation: Fan, X.; He, G.; Zhang, W.; Long, T.; Zhang, X.; Wang, G.; Sun, G.; Zhou, H.; Shang, Z.; Tian, D.; et al. Sentinel-2 Images Based Modeling of Grassland Above-Ground Biomass Using Random Forest Algorithm: A Case Study on the Tibetan Plateau. *Remote Sens.* **2022**, *14*, 5321. https://doi.org/10.3390/rs14215321

Academic Editor: Jun Ma

Received: 14 September 2022
Accepted: 21 October 2022
Published: 24 October 2022

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

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Grassland above-ground biomass (AGB) presents the total amount of organic matter produced by photosynthesis of grass per unit area [1,2]. AGB, which indicates energy flow and material circulation in the grassland ecosystem, is a critical indicator for grassland growing state, management and protection, playing a crucial role in carbon cycling and biodiversity conservation [3]. Hence, it is necessary to seek an appropriate method to estimate the grassland AGB of the Tibetan Plateau.

Currently, the methods of grassland biomass accounting contain field survey and remote sensing monitoring. The traditional field survey approach can obtain high-precision AGB data. At the same time, it requires a large amount of workforce, finance and time, and it is impossible to simultaneously assess the AGB over a vast area. Remote sensing monitoring is another option that has been widely used in AGB estimation at present. It has the advantage of a fast response time and being able to observe a large area of the land surface at the same time. Most previous studies on grassland AGB of the Tibetan Plateau used coarse spatial resolution images, such as MODIS, with a resolution of 250 m or 500 m [4–6]. However, coarse resolution remote sensing data ignore many surface details, particularly on the Tibetan Plateau, where the grassland is highly heterogeneous and fragmented, limiting the accuracy of ABG estimation. Therefore, higher resolution remote sensing data are needed to improve this situation.

The Copernicus Sentinel-2 (S2) mission comprises a constellation of two high spatial resolution multispectral imaging satellites. It has a spatial resolution of 10 m to 60 m and a revisit time of ten days with one satellite or five days for two satellites. Punalekar et al. [7] used S2A to estimate pasture Leaf Area Index and biomass on three sites in Southern England. Forkuor et al. [8] used multitemporal S2 data acquired over five months to map the AGB in the Sudanian Savanna of West Africa, achieving relatively high accuracy (R^2 = 0.83, RMSE = 60.6 Mg/ha). Compared with the 500 m resolution MODIS data, often used in previous studies, the higher resolution S2 images have more advantages in reducing the mixed pixels and decreasing the scale effect of remote sensing. However, the higher resolution of remote sensing images means that their data volume will also be more huge. This makes it difficult to use S2 data for mapping on large scales because the dramatic increase in data volume requires more storage space and more efficient methods of computation.

In remote-sensing-based AGB estimation, multiple features—such as spectral bands, vegetation indices (VI) and meteorological data—can be used as independent predictive variables for modeling [9,10]. Feature selection can select the optimal feature set among all the features to save computing resources, reduce feature space dimensions and improve the intelligibility and generalization of the model [11,12]. The Pearson correlation is a traditional method to understand the relationship between features and dependent variables, which measures the linear correlation between variables. An obvious drawback of the Pearson correlation coefficient is that it is sensitive only to linear relationships and does not reflect nonlinear ones. The mean impact value (MIV) increases one independent feature of the training set by 10% and then decreases it by 10% to obtain two new training sets. Then, the trained neural network is used to predict the two sets of results. The average of the two differences between the two predicted outcomes is the MIV of the features. This approach is more applicable to neural networks. Recursive Feature Elimination (RFE) is used to select relevant features in a training dataset for the prediction of target variable, which is an algorithm frequently applied in many machine learning approaches [13,14]. The RFE calculates the feature importance of each feature. Then, the feature set is recursively pruned to find the optimal set of features. RFE can effectively select the features in the training set that are relevant for predicting the target variable. Yin et al. [15] selected seven variables from eleven to estimate the height of the grassland by using the RFE method, increased the R^2 of the model from 0.41 to 0.51 and reduced the RMSE from 6.76 to 6.16 cm.

In the choice of modeling algorithms for inversion, machine learning algorithms are frequently used due to their ability to efficiently invert nonlinear relationships between variables [16–18]. Zhao et al. [19,20] compared seven different algorithms (including Cubist, RF, XGBoost, etc.) to determine the best model for DSM of clay content in topsoil and subsoil in semi-arid Australia. Yu et al. [21] showed the applications of the Gradient Boosting Regression Tree (GBRT), random forest (RF) and eXtremely Randomized Tree (ERT) in building AGB estimation models. Gao et al. [6] used 1200 AGB observations, NDVI, grassland types, latitude, longitude and altitude to develop an RF model suitable for AGB estimation in Tibetan alpine grasslands. However, little research has been conducted to integrate multiple features and compare different algorithms to estimate grassland AGB.

The Tibetan Plateau is the youngest, the largest and the highest plateau in the world. As the "Roof of the World", the average height of the Tibetan Plateau is above 4000 m [22]. The Tibetan Plateau and its surrounding mountains serve as the Third Pole in the northern hemisphere and is a climate-sensitive key ecological function area, acting as both the engine and amplifier of global change, like the Arctic Pole and the Antarctic Pole [23]. Acting as the primary land cover type of the Tibetan Plateau, grassland covers about 60% of the region area. Therefore, the Tibetan Plateau was selected as the case study for this research. The main goal of this study is to realize high-resolution AGB estimation using S2 as the main data source and to research the key techniques for model building and optimal feature selection. The major objectives were (1) to identify the significant features and find the optimal feature set in estimating AGB using S2; (2) to compare the performance of the Cubist, GBRT, RF and XGBoost algorithms in AGB estimation; and (3) to build an AGB estimation model using the best algorithm and analyze the uncertainty of the AGB estimation model.

2. Dataset

2.1. Study Area

The Tibetan Plateau (25°57′00″N∼39°48′00″N, 73°26′24″E∼104°25′12″E) is located in central Asia, and most of it is in China. The Tibetan Plateau has a length of approximately 2800 km from east to west and a width of 1000 km from north to west, while it has an area of about 2.5 million square kilometers [24]. The average altitude of the Tibetan plateau is over 4000 m. The air is dry and thin, with strong solar radiation, low temperature, and low rainfall. The average annual temperature is about 0.78 °C and the average annual precipitation is about 500 mm [25]. The whole study area is located in the alpine climate zone. The Tibetan Plateau is a sensitive area for global climate change. According to the land cover types map over the Tibetan Plateau [26], the grassland coverage of the Tibetan Plateau is 1.4×10^6 km^2, accounting for 53.65% of the total area, which is the main vegetation type of the Tibetan Plateau. The grassland types of the Tibetan Plateau include alpine meadow, alpine steppe and alpine desert, as shown in Figure 1.

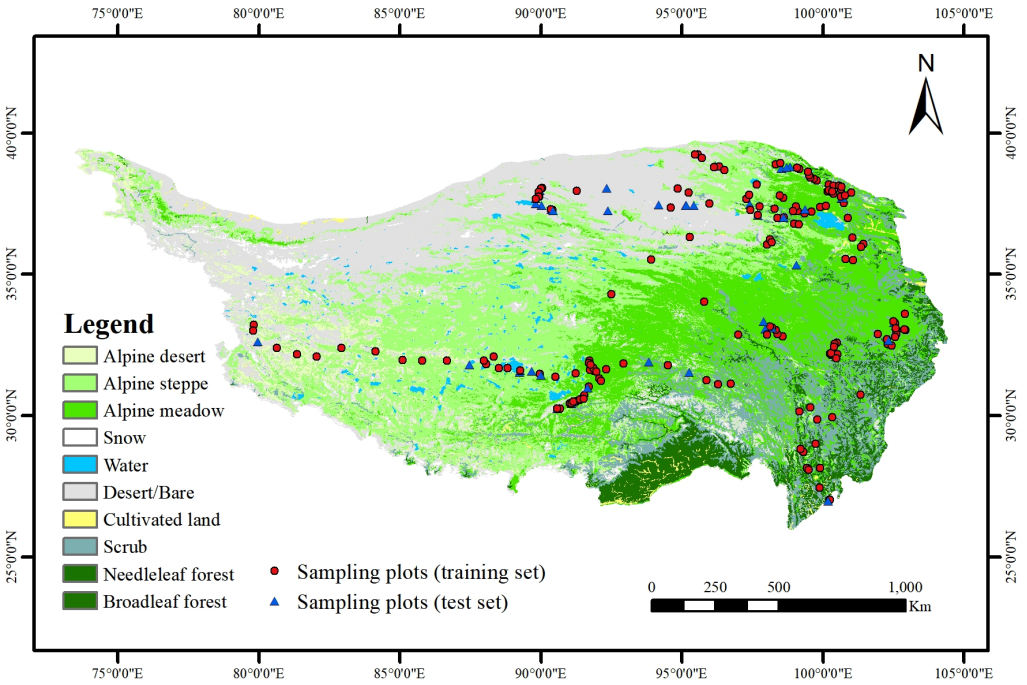

Figure 1. Spatial distribution of alpine grasses and AGB sampling plots on the Tibetan Plateau. The base map is the Alpine Grassland Map on the Tibetan Plateau [27].

2.2. Field Data

The field survey was carried out during July to September from 2019 to 2021, which is the growing season in the Tibetan Plateau. The sampling plots were mainly located on the Qiangtang Plateau, the Cocosili region, the southern slope of the Qilian Mountains, the Qaidam Basin, the Three-River Headwaters Region and the northern Sichuan–Tibet route (National Highway G317). The sampling points were sufficiently distributed over all types of grassland to ensure that the whole samples can represent the study area. The sampling plots were established with two 100 m red lines perpendicular to each other based on the center points, and the coordinates of the center points were recorded with a handheld GPS. We marked 10 m, 30 m and 50 m from the center point to form a box at 10 × 10 m, 30 × 30 m and 50 × 50 m scales. Aerial photographs were taken at these three scales using a drone as a ground identification reference. Each sample plot contained three (for homogeneous grass) to nine (for inhomogeneous grass) evenly distributed quadrats (1 m × 1 m). All plants from the quadrat were harvested and brought back to the laboratory for drying using an electric drying oven. The temperature of the drying oven was set to 65 °C. The plants were continuously dried in the drying oven until their weights no longer changed. Finally, the dried plants were weighed to obtain the dry matter (AGB). The data of sample plots consisted of grass type, coordinate information, elevation and dry weight. According to the land cover types map of the Tibetan Plateau, a total of 218 sampling plots of the grassland ecosystem were generated.

2.3. Elevation Dataset

The digital elevation model (DEM) data were obtained from the Shuttle Radar Topography Mission (SRTM) digital elevation dataset (version 4) (http://srtm.csi.cgiar.org (accessed on 14 September 2022)), which has a spatial resolution of 30 m. We calculated the slope and the angle data because topographic factors may also affect the growth of vegetation [28]. We extracted the pixel values of elevation, slope and aspect corresponding to each sample point on GEE.

2.4. Meteorological Data

In this study, we used the Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces dataset—containing Palmer Drought Severity Index (PDSI), soil moisture (Soil), Precipitation accumulation (Pr), Minimum temperature (tmmn) and Maximum temperature (tmmx)—provided by the University of California Merced (https://www.climatologylab.org/terraclimate.html (accessed on 14 September 2022)). This product covers 1958 to 2021, with a spatial and temporal resolution of 4638.3 meters and one month. We selected PDSI, Soil and Pr from the dataset that coincided with the month the ground survey was conducted as alternative input features. We obtained the monthly average temperature (T), which was also used as one of the alternative input features, by calculating the mean values of monthly tmmn and tmmx.

2.5. S2 Bands and S2-Derived VIs

We obtained the Sentinel-2 MultiSpectral Instrument Level-2A product (S2) from GEE (https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/product-types/level-2a (accessed on 14 September 2022)). Only those pixels that covered the sampling plots and were imaged closest to the survey date would be selected. If clouds contaminate a pixel on the day of the survey, then the band value for the survey day is replaced by the band value for the most adjacent date. In this study, the aerosol and the water vapor bands are considered irrelevant for AGB estimation and are therefore not used. The remaining ten bands were selected as features to build the grassland AGB estimation models, with resolutions of 10 m × 10 m and 20 m × 20 m (Table 1). Besides, the Difference Vegetation Index (DVI), Enhanced Vegetation Index (EVI), Modified Soil Adjusted Vegetation Index (MSAVI), Modified Simple Ratio (MSR), Normalized Difference Infrared Index (NDII), Normalized Difference Phenology Index (NDPI), Normalized Difference Vegetation Index

(NDVI), Optimized Soil-Adjusted Vegetation Index (OSAVI), Ratio Vegetation Index (RVI) and Soil-Adjusted Vegetation Index (SAVI) were selected as the alternative features as well. As these 10 VIs were derived from the spectral reflectance data, they can reflect vegetation growth, physiological characteristics, growing environment and reduction of soil background effects. The calculation of these VIs can be referred to in Table 2.

Table 1. Information about the Sentinel-2 bands selected in this study.

Sentinel-2 Bands	Central Wavelength (μm)	Resolution (m)	Wavelength S2A/S2B (nm)
B2—Blue	0.490	10	496.6/492.1
B3—Green	0.560	10	560/559
B4—Red	0.665	10	664.5/665
B5—Red Edge 1	0.705	20	703.9/703.8
B6—Red Edge 2	0.740	20	740.2/739.1
B7—Red Edge 3	0.783	20	782.5/779.7
B8—NIR	0.842	10	835.1/833
B8A—Narrow NIR	0.865	20	864.8/864
B11—SWIR 1	1.610	20	1613.7/1610.4
B12—SWIR 2	2.190	20	2202.4/2185.7

Table 2. Vegetation indices used in the AGB estimation model.

Vegetation Index	Formula	References
DVI	$NIR - RED$	[29,30]
EVI	$G \times \frac{NIR-RED}{NIR + C_1 \times RED - C_2 \times BLUE + L}$ *where* $G = 2.5; C_1 = 6; C_2 = 7.5; L = 1$	[31,32]
MSAVI	$\frac{1}{2} \times \left[(2 \times NIR + 1) - \left(\sqrt{(2 \times NIR + 1)^2 - 8 \times (NIR - RED)}\right)\right]$	[33]
MSR	$\frac{\frac{NIR}{RED} - 1}{\sqrt{\frac{NIR}{RED}} + 1}$	[34]
NDII	$\frac{NIR-SWIR}{NIR + SWIR}$	[35]
NDPI	$\frac{NIR-(0.74 \times RED + 0.26 \times SWIR)}{NIR + (0.74 \times RED + 0.26 \times SWIR)}$	[36]
NDVI	$\frac{NIR-RED}{NIR + RED}$	[37,38]
OSAVI	$\frac{NIR-RED}{NIR + RED + X}$ where $X = 0.16$	[39,40]
RVI	$\frac{NIR}{RED}$	[41]
SAVI	$\frac{NIR-RED}{NIR + RED + L} \times (1 + L)$ *where* $L = 0.5$	[42,43]

3. Methodology

3.1. AGB Estimation Procedure

Figure 2 shows the procedure of establishing the AGB estimation model and the evaluation analysis. The first step was preparing the field, meteorological, topographic and spectral data (including S2 bands and vegetation indices). Then, they were divided into training and test sets according to model building needs and spatial–temporal scalability analysis. The second step was calculating the feature importance and selecting the optimal feature set using the RFE algorithm. The third step was to build the RF model based on the optimal feature set to estimate the grassland AGB and then recalculate the model's accuracy

by inputting the training and test sets with specific spatial and temporal segmentation, respectively. The final step was to analyze the spatial–temporal scalability and uncertainty of the model.

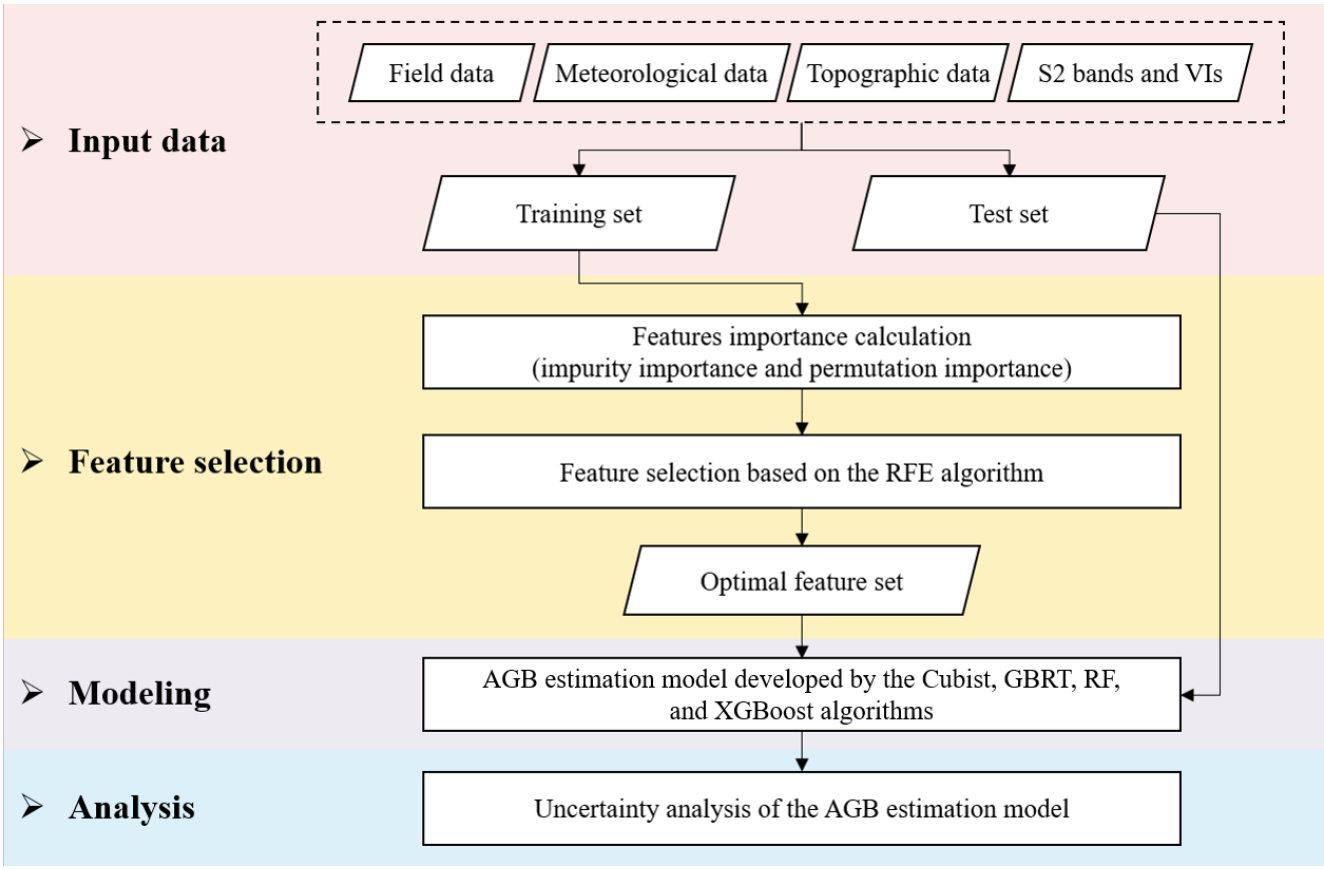

Figure 2. Flowchart of the grassland AGB estimation and analysis.

3.2. Recursive Feature Elimination

RFE is a feature selection method that considers smaller feature sets by recursion. First, the initial feature set is trained to estimate the importance of each feature obtained by any specific algorithm. In this study, two algorithms—impurity and permutation importance—were used to calculate the feature importance (see Section 3.3 for details). Then, the feature with the lowest feature importance score is deleted from the current feature set. This process is repeated recursively on the pruned set. The calculation stops when the feature set reaches a predefined number of features. The procedure of the RFE algorithm is shown in Figure 3.

3.3. Feature Importance

Feature importance is the extent to which a feature is relevant to the target variable. It can be calculated in several ways. The RF algorithm has built-in feature importance computed by the index "Gini importance". It is defined as the total impurity reduction of all nodes averaged over all ensemble trees [44,45]. The biggest advantage of this method is that all feature importance can be calculated during the random forest training. The disadvantage of this method is that it tends to select numerical features with high cardinality. Moreover, it selects one of the features and ignores the importance of the second in the case of having relevant features, which may lead to the wrong conclusions.

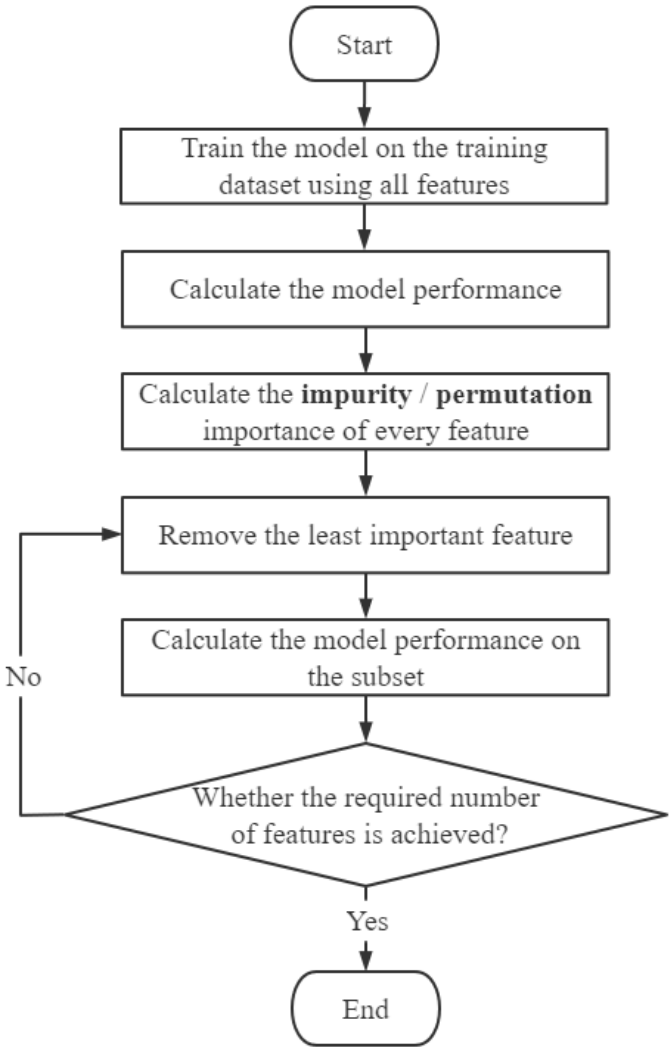

Figure 3. The procedure of the RFE algorithm.

Permutation-based importance [46,47] can override the drawbacks of default feature importance calculated by the mean decrease in node impurity. The calculation steps of permutation importance are as follows: First, an initial score is evaluated on a dataset consisting of the original features. Next, the scores are evaluated again by randomly corrupting the value of a feature and permuting the corresponding feature columns on the test set, keeping the target variable and other features unchanged. The difference between the initial score and the scores in the permutating feature columns is defined as the permutation importance. In this study, we weighted the permutation importance of all features so that their permutation importance would add up to 1 to make the impurity importance and the permutation importance comparable. We repeated the above two methods 1000 times to calculate the average values as the ultimate feature importance scores.

3.4. Cubist

Cubist is a rule-based regression tree algorithm and an extension of the M5 model tree [48]. The algorithm generates rule-based models with one or more rules, each rule containing a set of criteria associated with a multivariate linear submodel. Each linear model is a "leaf" of Cubist. Therefore, the Cubist model is efficient and easy to understand. The number of rules was set to 500 in this study.

3.5. GBRT

The Gradient Boosting Regression Tree (GBRT) is an ensemble learning model. This algorithm combines multiple regression trees to generate a gradient-enhanced estimation model [49]. These regression trees are arranged in a string. The GBRT is designed to first fit the data with the first regression tree and calculate the residuals between the fitted results and the true values. The second regression tree then continues to fit the residuals in the previous step to reduce the residuals between the overall fitted values and the true values. The number of iterations depends on the number of regression trees. The number of estimators was set to 1000 and the learning rate was set to 0.1.

3.6. RF

The random forest (RF) regression algorithm belongs to the bootstrap aggregation method of an ensemble learning algorithm in which multiple random regression trees are combined to achieve better regression accuracy [50]. These regression trees randomly extract data from the training set and vote equally to obtain an average as the final result. The regression trees in the RF are parallel and constructed by random vectors that are sampled independently. The RF is widely used in AGB estimation due to its advantages of processing high-dimensional vectors, reducing over-fitting, fast training speed and noise immunity to a certain extent. The result of the model is obtained by averaging the results of all regression trees. Theoretically, the larger the number of regression trees, the lower the impact of extreme cases on the true results. Therefore, we set the number of regression trees to 1000, and each split had at least two variables [51,52].

3.7. XGBoost

The eXtreme Gradient Boosting (XGBoost) is an extension of GBRT and also an ensemble learning model by stringing multiple regression trees. In contrast to GBRT, XGBoost introduces a second-order Taylor formula, while GBRT uses first-order derivatives. In addition, XGBoost can correct the residuals and create a new tree based on the previous tree [53]. The number of estimators was set to 1000 and the learning rate was set to 0.055.

3.8. Estimation Accuracy Evaluation

The accuracy of the model was assessed by using independently observed AGB field measurements. The entire sampling plots were randomly separated into two sets without any overlap, with one set serving as the training set and the other as the test set. The sample plots used as the training and test set were distributed in all months and regions. Model evaluation was performed using regression statistics, including coefficient of determination (R^2), root mean square error (RMSE), Lin's concordance correlation coefficient (LCCC) and the ratio of performance to deviation (RPD).

$$R^2 = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(y_i - \overline{y}_i)^2 \quad (1)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / n} \quad (2)$$

where y_i is the observed AGB value, $\hat{y}_i$ is the predicted AGB value, y is the arithmetic mean of all the observed AGB values and n is the number of samples in the training or test set. In general, a higher R^2 value and a lower RMSE value indicate that the model has better estimation performance.

The LCCC determines the predicted distance on a 45-degree line from the origin to the measured data.

$$LCCC = \frac{2 s_x y}{s_x^2 + s_y^2 + (\overline{x} - \overline{y})^2} \quad (3)$$

$$s_x y = \frac{1}{N} \sum_{n=1}^{N} (x_n - \overline{x})(y_n - \overline{y}) \quad (4)$$

where $\overline{x}$ and $\overline{y}$ are the means for the predicted and observed AGB values, s_x^2 and s_y^2 are the corresponding variances and N is the number of samples in the validation dataset. The level of predicted agreement was evaluated according to Rossel et al. [54], where LCCC = 1 indicates perfect agreement. An LCCC value > 0.9 indicates excellent agreement, while a value between 0.80 and 0.90 indicates good agreement. An LCCC value between 0.65 and 0.80 indicates moderate agreement, and an LCCC < 0.65 indicates poor agreement.

Prediction accuracy is assessed using the ratio of performance to deviation (RPD), which is calculated as the ratio of the standard deviation (SD) to the RMSE, as follows:

$$RPD = \frac{SD}{RMSE} \quad (5)$$

According to Rossel et al. [55], the RPD can be graded as excellent (>2.5), very good (2.0–2.5), good (1.8–2.0), fair (1.4–1.8) and poor (<1.4) prediction accuracy.

4. Results and Discussion

4.1. Spatial Distribution Grassland AGB from All the Sampling Plots

Figure 4 illustrates the spatial distribution of the measured grassland AGB from 2019 to 2021 on the Tibetan Plateau. The AGB values range from 6.47 g/m² to 602.23 g/m², the median value is 59.46 g/m² and the average value is 97.41 g/m². There were 96 (44.04%) sampling plots with AGB values less than 50 g/m² and 149 (68.35%) sampling plots less than 100 g/m². The sampling plots were scattered in the northeastern and southwestern parts of the Tibetan Plateau, with the Qaidam Basin and the southern slopes of the Qilian Mountains in the northeast and the sampling plots in the southwest, mainly along the National Highway G317. In the northwestern part of the Tibetan Plateau are the Hoh Xil region and the Qiangtang Plateau, where the harsh environment makes it difficult to conduct ground surveys, and the main land cover type in the southwestern part is forest, with relatively few sampling plots. The high-value sampling plots of AGB are mainly concentrated in the eastern part of the Tibetan Plateau, and the value of AGB gradually decreases from east to west.

4.2. Feature Importance

The feature importance scores were investigated by the total decrease in node impurities and the mean decrease in permutation accuracy (Figure 5). The distributions of feature importance scores calculated by the impurity and permutation importance for RF models were not similar. Feature importance scores based on the permutation importance are more concentrated on a few features, with the top five contributing more than 60%. In contrast, feature importance scores based on impurity importance are more evenly distributed, with eight features required to achieve the same contribution. However, the importance score of the temperature variable ranked first in both feature importance methods, implying that the grassland AGB estimation may be highly sensitive to temperature. Additionally, two feature variables related to canopy water content, NDPI and NDII, also ranked high in both feature importance calculation methods (ranked 2nd and 3rd for impurity importance, ranked 3rd and 4th for permutation importance, respectively).

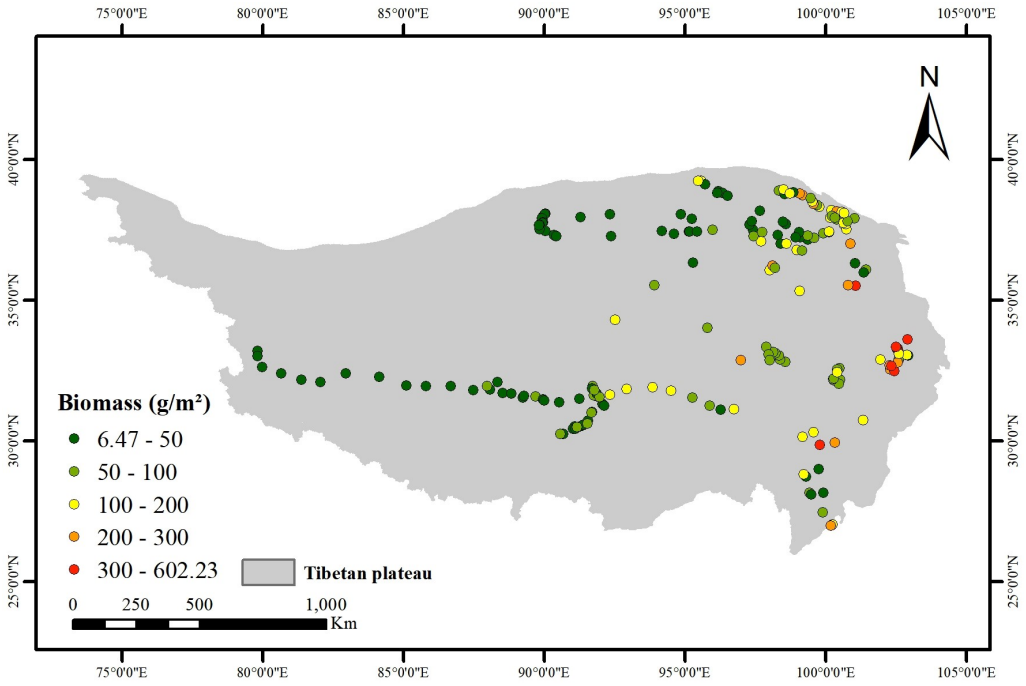

Figure 4. The spatial distribution of the measured grassland AGB values on the Tibetan Plateau during 2019 to 2021, including 218 sampling plots.

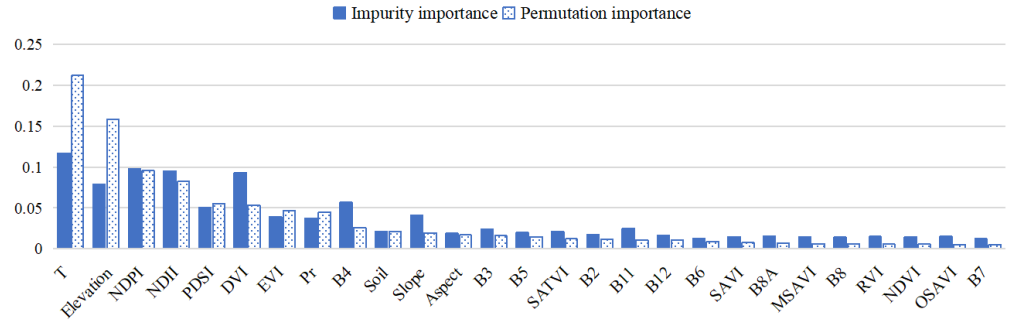

Figure 5. The feature importance scores based on impurity importance and permutation importance.

4.3. Performance of Feature Selection

The RFE algorithm based on impurity importance and permutation importance was used in the feature selection. Figure 6 illustrates how the R^2 values and RMSE values vary with the number of selected features of RF models on the training and test sets. Each line represents an individual model's R^2, and the bar at the bottom of the chart is its corresponding RMSE.

For the training set, the models' accuracies based on the two feature importance methods varied little, with R^2 ranging from a minimum of 0.86 to a maximum of 0.93, and RMSE from a maximum of 36.97 g/m^2 to a minimum of 25.27 g/m^2. The accuracy of the models on the training set gradually rose and eventually saturated as the number of features increased. For the test set, the results of the two models differed. The accuracy of the impurity importance-based model remarkably increased until it reached two features and then gradually stabilized and reached the maximum accuracy when the number of features reached 10. The accuracy of the permutation importance-based model increased incrementally with the number of features until the maximum accuracy was reached at five features. Then, the accuracy decreased slightly but the R^2 was still above 0.8.

Comparing the two feature importance algorithms, the permutation importance-based model achieved the highest accuracy with fewer input features (5 features); so, it was superior to the impurity importance-based model in this study.

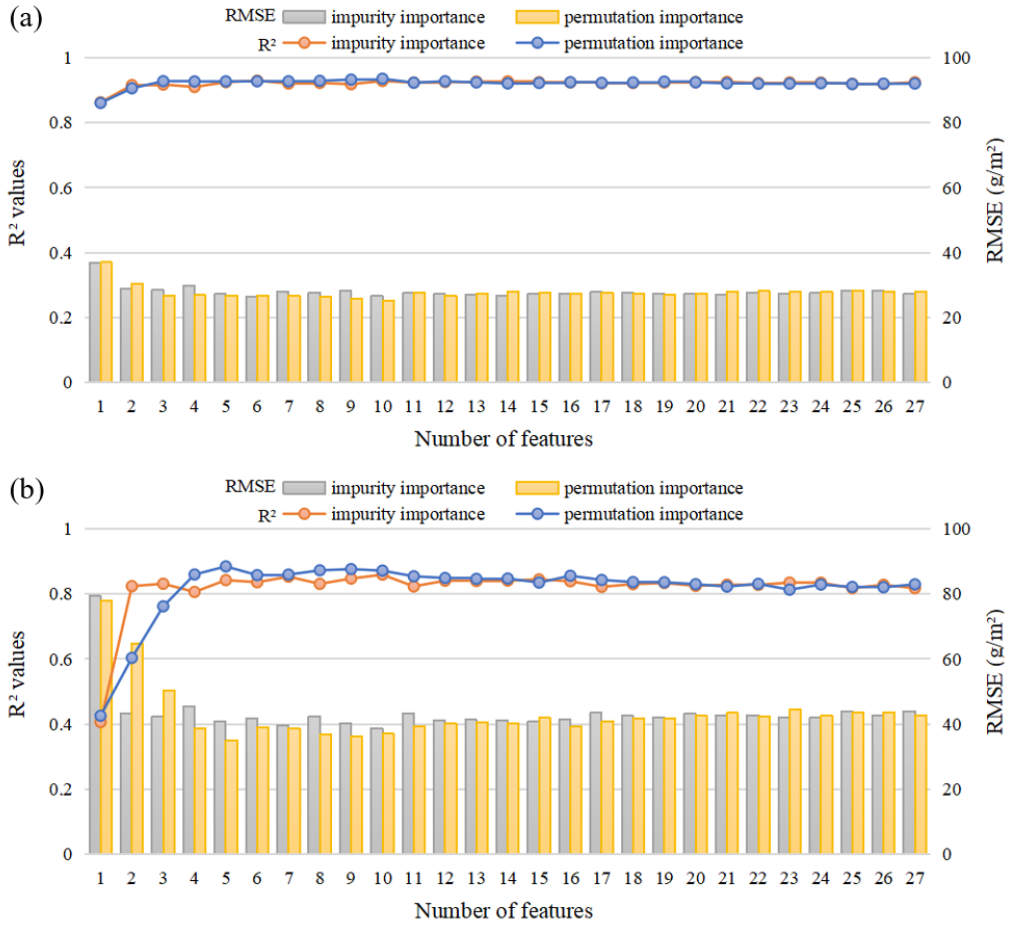

Figure 6. The performance of the AGB estimation models with the selected feature number changes based on the impurity and permutation importance. The left vertical axes (curves) and right vertical axes (bars) indicate the R^2 and RMSE of the estimation models. (**a**,**b**) show the model performance on the training and test sets, respectively.

4.4. Evaluation of AGB Models

Two RF models with the highest accuracy based on the impurity and the permutation importance were obtained by the RFE algorithm. For impurity importance, model accuracy reached a maximum of 0.8585 when the number of features reached 10 (T, NDPI, NDII, DVI, elevation, B4, PDSI, slope, EVI and Pr). For permutation importance, the maximum model accuracy value is 0.8838 when the number of features is 5 (T, elevation, NDPI, NDII and PDSI). Therefore, we choose the features based on the permutation importance as the input parameters of the models.

The performance of the Cubist, RF, GBRT and XGBoost models on the test set was expressed as scatter plots that showed the relationship between the estimated AGB values and the observed AGB values (Figure 7). As shown in the figure, the RF model (Figure 7c: R^2 = 0.8838, RMSE = 35.05 g/m^2, LCCC = 2.44, RPD = 0.91) performs better than the Cubist (Figure 7a: R^2 = 0.6221, RMSE = 63.22 g/m^2, LCCC = 1.06, RPD = 0.72), the GBDT (Figure 7b: R^2 = 0.8318, RMSE = 42.17 g/m^2, LCCC = 2.18, RPD = 0.89) and the XGBoost (Figure 7d: R^2 = 0.8272, RMSE = 42.74 g/m^2, LCCC = 1.99, RPD = 0.88) model. It is worth noting that the underestimation of AGB in the high-value range and the overestimation of AGB in the low-value range by both models are still inevitable. Therefore, the RF model based on T, elevation, NDPI, NDII and PDSI is considered to be the optimal model for estimating the grassland AGB of the Tibetan Plateau.

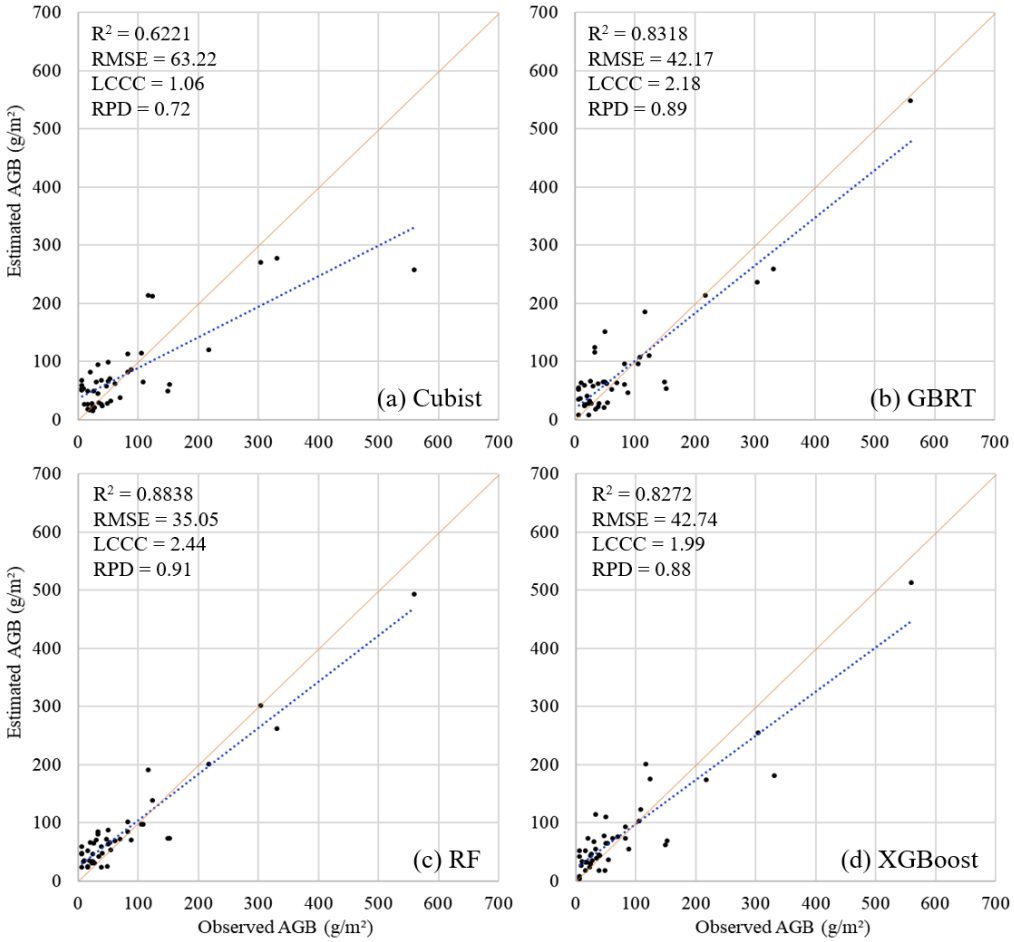

Figure 7. Scatter plots of the estimated and observed AGB based on (**a**) the Cubist, (**b**) the GBDT, (**c**) RF and (**d**) the XGBoost model using the test set.

4.5. Advantages of S2 Images and S2-Derived VIs in Estimating Grassland AGB

The implicit premise of applying remote sensing to invert surface parameters is that different ground objects or the same object in different growth states have different spectral characteristics. Specifically, they may have different reflectance in different wavelength bands. The spatial resolution of MODIS data used in previous studies is coarse, with a resolution of 500 m. There is a high potential for mixed image pixels, where different types of objects exist within a single image pixel, in the Tibetan Plateau region, where the ground cover is fragmented, and the grassland heterogeneity is large. In this study, we used S2 images with a spatial resolution of 10 m (bands 2, 3, 4 and 8) and 20 m (bands 5, 6, 7, 8A, 11 and 12) to reduce the negative impact of mixed pixels.

Second, the spatial complexity of features is often reflected in the spatial structure of features, the radiometric properties of features themselves and the differences in vegetation components, resulting in quantitative remote sensing products with scale effects. The spatial scale effect in quantitative remote sensing science is defined as the same area, the same time, the same remote sensing model, the same kind of remote sensing data and the same imaging conditions but the difference in resolution leads to inconsistency in remote sensing inversion surface parameters. In the ground survey experiment, the sample size is 1 m × 1 m, which is much smaller than the size of a MODIS image pixel (500 m × 500 m). The pixel size of the S2 used in this experiment is 10 m × 10 m and 20 m × 20 m, which is still larger than the area of one sample. Still, it greatly improved the problem of spatial heterogeneity differences and scale effects between sample data and satellite data. Our results showed that the estimation model based on the S2 images had an accuracy of 0.8838, which indicates that S2 images have an advantage in estimating grassland AGB.

4.6. Influence of Features on Grassland AGB

T, elevation, NDPI, NDII and PDSI were eventually selected as input features for the RF model. NDPI has two major advantages in the estimation of grassland AGB [36]. First, NDPI is highly sensitive to vegetation, and many types of soil have NDPI values close to zero. Therefore, NDPI can alleviate the negative effects of heterogeneous soil background on AGB estimation in grasslands. Second, NDPI contains the SWIR band. Water absorption is significantly increased in SWIR. Therefore, NDPI is sensitive to leaf water content and can capture the change in leaf water content [36]. NDII is a vegetation index that is sensitive to changes in plant canopy water content. It is also a valid indicator of water storage in the root zone of vegetation during water deficit and a strong indicator for assessing drought [35]. The values of NDII increase with increasing water content. PDSI is based on the principle of soil water balance and is used to characterize the deficit of actual soil water supply in an area at a certain time relative to the local climate suitable water supply [56]. Precipitation is usually used to replace the water supply, and the calculation of water demand involves evapotranspiration and soil moisture changes. The selection of NDPI, NDII and PDSI implies that grassland AGB is significantly influenced by leaf water content and drought conditions, which is consistent with the results of [6,57]. T represents the monthly average temperature. The feature of elevation has a high feature importance score, indicating that there are obvious differences in the AGB vertical distribution on the grasslands of the Tibetan Plateau. In summary, temperature, altitude, humidity and leaf water content greatly affect grassland AGB, and features associated with these factors play an important role in estimating grassland AGB.

4.7. Limitations and Future Works

This study shows that although the RF model based on T, elevation, NDPI, NDII and PDSI performs well in improving the accuracy of grassland AGB estimation model (R^2 = 0.8838, RMSE = 35.05 g/m^2, LCCC = 2.44, RPD = 0.91), the accuracy of the model is still influenced by a few factors. First, the dates of the field measurements may not match the dates of the remotely sensed images. S2 products have a temporal resolution of 5 days, and some of these image products cannot be used due to cloud cover. Although we selected the image nearest to the field measurement date, it may still not reflect the true AGB of the sampling date, even though the change in AGB between the sampling date and the date of the S2 product may not be significant. In the future, we will consider fusing multisource remote sensing images to reduce the discrepancy between time phases.

Second, the RF model used in this study is a data-driven ensemble learning model. The data-driven model is very sensitive to data, and its performance usually depends on the quantity and distribution of data. Due to the harsh natural conditions and limited transportation conditions, the number of sampling plots distributed in the western Tibetan Plateau is significantly less than that in the densely populated eastern region. As a result, the model's training data are uneven in the spatial distribution of the Tibetan Plateau. The model learns prior information about the proportion of samples in the training set so that it may pay more attention to the regions with a larger number of sampling plots in the actual prediction. The uneven spatial distribution of the sample plots affects the model's ability to learn essential mapping relations between AGB and input features and reduces the model's spatial scalability.

Third, different grassland types have diverse characteristics, and the data used in this study do not contain the feature of grassland types as the current amount of sampling data in this study is not enough to support modeling according to different grassland types. We had a total of 218 available data, and after grouping by different grassland types, 48, 58 and 112 data were available for alpine desert, alpine grassland and alpine meadow, respectively. Since the random forest algorithm is a data-driven algorithm, the amount of data is very important to the results. Only by ensuring that the amount of data is sufficient can we ensure that the model results are stable and reliable. In addition, the field measurements in this study were all from the growing season of the grassland. Lack of data reduced the temporal scalability of the model and limited the model's accuracy in estimating AGB

in nongrowing seasons. In future studies, we will conduct ground surveys and expand the number of sample plots in the regions with sparse sampling plots and during the nongrowing season to improve the model accuracy.

Besides, the highly complex topographic factors of the Tibetan Plateau, such as its varied slope and aspect, have an effect on the reflectance of the grassland, which may affect the data received by the remote sensor and make the values of vegetation index biased. These errors would eventually affect the accuracy of the AGB model. In future work, we will strive to integrate machine learning algorithms and physical models or process-based biogeochemical models [58], enhance model interpretation and reduce the influence of terrain factors to build better models for AGB estimation.

5. Conclusions

In this study, we selected 10 S2 bands, 10 S2-derived VIs, 218 pieces of AGB field survey data, four types of meteorological data and three types of topographic data as the alternative input features of the model. The RFE algorithm based on two feature importance algorithms (impurity importance and permutation importance) was used to find the optimal feature set. Then, we compared the performance of the Cubist, GBRT, RF and XGBoost algorithms. The results showed that the RF model based on the feature set of T, elevation, NDPI, NDII and PDSI performed best (R^2 = 0.8838, RMSE = 35.05 g/m^2, LCCC = 2.44, RPD = 0.91). The results of feature selection indicate that features associated with temperature, altitude, humidity and leaf water content significantly influence grassland AGB. To achieve more accurate estimation results, more sampling plots need to be collected in the sparse area and during the no-growing seasons. Moreover, physical models and process-based biogeochemical models can be introduced to reduce the impact of terrain.

Author Contributions: Methodology, X.F., G.H. and W.Z.; AGB field data, T.L., X.Z., G.W., G.S., H.Z., Z.S., D.T., X.L. and X.S.; experiment, X.F.; results analysis, X.F. and G.H.; data curation, X.F.; writing—original draft preparation, X.F.; writing—review and editing, X.F., G.H. and T.L.; project administration, G.H.; funding acquisition, G.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Second Tibetan Plateau Scientific Expedition and Research Program (STEP), grant number 2019QZKK030701; the program of the National Natural Science Foundation of China, grant number 61731022; and the Strategic Priority Research Program of the Chinese Academy of Sciences, grant number XDA19090300.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank the anonymous reviewers and the editors for their valuable comments to improve our manuscript.

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

References

1. Nie, X.-Q.; Yang, L.-C.; Xiong, F.; Li, C.-B.; Fan, L.; Zhou, G.-Y. Aboveground biomass of the alpine shrub ecosystems in three-river source region of the tibetan plateau. *J. Mt. Sci.* **2018**, *15*, 57–363. [CrossRef]
2. John, R.; Chen, J.; Giannico, V.; Park, H.; Xiao, J.; Shirkey, G.; Ouyang, Z.; Shao, C.; Lafortezza, R.; Qi, J. Grassland canopy cover and aboveground biomass in mongolia and inner mongolia: Spatiotemporal estimates and controlling factors. *Remote Sens. Environ.* **2018**, *213*, 34–48. [CrossRef]
3. Xia, J.; Ma, M.; Liang, T.; Wu, C.; Yang, Y.; Zhang, L.; Zhang, Y.; Yuan, W. Estimates of grassland biomass and turnover time on the tibetan plateau. *Environ. Res. Lett.* **2018**, *13*, 014020. [CrossRef]
4. Liu, S.; Cheng, F.; Dong, S.; Zhao, H.; Hou, X.; Wu, X. Spatiotemporal dynamics of grassland aboveground biomass on the qinghai-tibet plateau based on validated modis ndvi. *Sci. Rep.* **2017**, *7*, 4182. [CrossRef] [PubMed]
5. Zeng, N.; Ren, X.; He, H.; Zhang, L.; Zhao, D.; Ge, R.; Li, P.; Niu, Z. Estimating grassland aboveground biomass on the tibetan plateau using a random forest algorithm. *Ecol. Indic.* **2019**, *102*, 479–487. [CrossRef]
6. Gao, X.; Dong, S.; Li, S.; Xu, Y.; Liu, S.; Zhao, H.; Yeomans, J.; Li, Y.; Shen, H.; Wu, S.; et al. Using the random forest model and validated modis with the field spectrometer measurement promote the accuracy of estimating aboveground biomass and coverage of alpine grasslands on the qinghai-tibetan plateau. *Ecol. Indic.* **2020**, *112*, 106114. [CrossRef]

7. Punalekar, S.M.; Verhoef, A.; Quaife, T.; Humphries, D.; Bermingham, L.; Reynolds, C. Application of sentinel-2a data for pasture biomass monitoring using a physically based radiative transfer model. *Remote Sens. Environ.* **2018**, *218*, 207–220. [CrossRef]
8. Forkuor, G.; Zoungrana, J.-B.B.; Dimobe, K.; Ouattara, B.; Vadrevu, K.P.; Tondoh, J.E. Above-ground biomass mapping in west african dryland forest using sentinel-1 and 2 datasets-a case study. *Remote Sens. Environ.* **2020**, *236*, 111496. [CrossRef]
9. Zhang, Y.; Shao, Z. Assessing of urban vegetation biomass in combination with lidar and high-resolution remote sensing images. *Int. J. Remote Sens.* **2021**, *42*, 964–985. [CrossRef]
10. Lu, D. Aboveground biomass estimation using landsat tm data in the brazilian amazon. *Int. J. Remote Sens.* **2005**, *26*, 2509–2525. [CrossRef]
11. Yu, K.; Wu, X.; Ding, W.; Pei, J. Scalable and accurate online feature selection for big data. *ACM Trans. Knowl. Discov. Data (TKDD)* **2016**, *11*, 1–39. [CrossRef]
12. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–45. [CrossRef]
13. Meyer, H.; Lehnert, L.W.; Wang, Y.; Reudenbach, C.; Nauss, T.; Bendix, J. From local spectral measurements to maps of vegetation cover and biomass on the qinghai-tibet-plateau: Do we need hyperspectral information? *Int. J. Appl. Earth Obs. Geoinf.* **2017**, *55*, 21–31. [CrossRef]
14. Misra, P.; Yadav, A.S. Improving the classification accuracy using recursive feature elimination with cross-validation. *Int. J. Emerg. Technol.* **2020**, *11*, 659–665.
15. Yin, J.; Feng, Q.; Liang, T.; Meng, B.; Yang, S.; Gao, J.; Ge, J.; Hou, M.; Liu, J.; Wang, W.; et al. Estimation of grassland height based on the random forest algorithm and remote sensing in the tibetan plateau. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *13*, 178–186. [CrossRef]
16. Haq, M.A. Planetscope nanosatellites image classification using machine learning. *Comput. Syst. Sci. Eng.* **2022**, *42*, 1031–1046.
17. Wang, X.; Zhang, J.; Xun, L.; Wang, J.; Wu, Z.; Henchiri, M.; Zhang, S.; Zhang, S.; Bai, Y.; Yang, S.; et al. Evaluating the effectiveness of machine learning and deep learning models combined time-series satellite data for multiple crop types classification over a large-scale region. *Remote Sens.* **2022**, *14*, 2341. [CrossRef]
18. Pelta, R.; Beeri, O.; Tarshish, R.; Shilo, T. Sentinel-1 to ndvi for agricultural fields using hyperlocal dynamic machine learning approach. *Remote Sens.* **2022**, *14*, 2600. [CrossRef]
19. Zhao, D.; Arshad, M.; Wang, J.; Triantafilis, J. Soil exchangeable cations estimation using vis-nir spectroscopy in different depths: Effects of multiple calibration models and spiking. *Comput. Electron. Agric.* **2021**, *182*, 105990. [CrossRef]
20. Zhao, D.; Wang, J.; Zhao, X.; Triantafilis, J. Clay content mapping and uncertainty estimation using weighted model averaging. *Catena* **2022**, *209*, 105791. [CrossRef]
21. Yu, R.; Yao, Y.; Wang, Q.; Wan, H.; Xie, Z.; Tang, W.; Zhang, Z.; Yang, J.; Shang, K.; Guo, X.; et al. Satellite-derived estimation of grassland aboveground biomass in the three-river headwaters region of china during 1982–2018. *Remote Sens.* **2021**, *13*, 2993. [CrossRef]
22. Tang, Y.; Wan, S.; He, J.; Zhao, X. Foreword to the special issue: Looking into the impacts of global warming from the roof of the world. *J. Plant Ecol.* **2009**, *2*, 169–171. [CrossRef]
23. You, Y.; Wang, S.; Pan, N.; Ma, Y.; Liu, W. Growth stage-dependent responses of carbon fixation process of alpine grasslands to climate change over the tibetan plateau, china. *Agric. For. Meteorol.* **2020**, *291*, 108085. [CrossRef]
24. Ding, M.; Zhang, Y.; Sun, X.; Liu, L.; Wang, Z.; Bai, W. Spatiotemporal variation in alpine grassland phenology in the qinghai-tibetan plateau from 1999 to 2009. *Chin. Sci. Bull.* **2013**, *58*, 396–405. [CrossRef]
25. Ding, M. Temperature and Precipitation Grid Data of the Qinghai Tibet Plateau and Its Surrounding Areas in 1998–2017. Grid Data of Annual Temperature and Annual Precipitation on the Tibetan Plateau and Its Surrounding Areas during 1998–2017. Available online: https://data.tpdc.ac.cn/en/data/c954daad-6086-4edd-a6c5-f69c581e5c31/ (accessed on 14 September 2022).
26. Wang, C.; Guo, H.; Zhang, L.; Qiu, Y.; Sun, Z.; Liao, J.; Liu, G.; Zhang, Y. Improved alpine grassland mapping in the tibetan plateau with modis time series: A phenology perspective. *Int. J. Digit. Earth* **2015**, *8*, 133–152. [CrossRef]
27. Wang, C.; Guo, H.; Zhang, L.; Qiu, Y.; Sun, Z.; Liao, J.; Liu, G.; Zhang, Y. Alpine Grassland Map. Science Data Bank. 2017. Available online: https://www.scidb.cn/en/detail?dataSetId=633694460949037059 (accessed on 14 September 2022).
28. Liang, T.; Yang, S.; Feng, Q.; Liu, B.; Zhang, R.; Huang, X.; Xie, H. Multi-factor modeling of above-ground biomass in alpine grassland: A case study in the three-river headwaters region, china. *Remote Sens. Environ.* **2016**, *186*, 164–172. [CrossRef]
29. Yan, F.; Wu, B.; Wang, Y. Estimating aboveground biomass in mu us sandy land using landsat spectral derived vegetation indices over the past 30 years. *J. Arid Land* **2013**, *5*, 521–530. [CrossRef]
30. Wang, G.; Liu, S.; Liu, T.; Fu, Z.; Yu, J.; Xue, B. Modelling above-ground biomass based on vegetation indexes: A modified approach for biomass estimation in semi-arid grasslands. *Int. J. Remote Sens.* **2019**, *40*, 3835–3854. [CrossRef]
31. Liu, H.Q.; Huete, A. A feedback based modification of the ndvi to minimize canopy background and atmospheric noise. *IEEE Trans. Geosci. Remote Sens.* **1995**, *33*, 457–465. [CrossRef]
32. Shen, M.; Tang, Y.; Klein, J.; Zhang, P.; Gu, S.; Shimono, A.; Chen, J. Estimation of aboveground biomass using in situ hyperspectral measurements in five major grassland ecosystems on the tibetan plateau. *J. Plant Ecol.* **2008**, *1*, 247–257. [CrossRef]
33. Qi, J.; Chehbouni, A.; Huete, A.R.; Kerr, Y.H.; Sorooshian, S. A modified soil adjusted vegetation index. *Remote Sens. Environ.* **1994**, *48*, 119–126. [CrossRef]
34. Chen, J.M. Evaluation of vegetation indices and a modified simple ratio for boreal applications. *Can. J. Remote Sens.* **1996**, *22*, 229–242. [CrossRef]

35. Sriwongsitanon, N.; Gao, H.; Savenije, H.; Maekan, E.; Saengsawan, S.; Thianpopirug, S. The normalized difference infrared index (ndii) as a proxy for soil moisture storage in hydrological modelling. *Hydrol. Earth Syst. Sci. Discuss.* **2015**, *12*, 8419–8457.
36. Xu, D.; Wang, C.; Chen, J.; Shen, M.; Shen, B.; Yan, R.; Li, Z.; Karnieli, A.; Chen, J.; Yan, Y.; et al. The superiority of the normalized difference phenology index (ndpi) for estimating grassland aboveground fresh biomass. *Remote Sens. Environ.* **2021**, *264*, 112578. [CrossRef]
37. Gamon, J.A.; Field, C.B.; Goulden, M.L.; Griffin, K.L.; Hartley, A.E.; Joel, G.; Penuelas, J.; Valentini, R. Relationships between NDVI, canopy structure, and photosynthesis in three Californian vegetation types. *Ecol. Appl.* **1995**, *5*, 28–41. [CrossRef]
38. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [CrossRef]
39. Rondeaux, G.; Steven, M.; Baret, F. Optimization of soil-adjusted vegetation indices. *Remote Sens. Environ.* **1996**, *55*, 95–107. [CrossRef]
40. Ren, H.; Zhou, G.; Zhang, F. Using negative soil adjustment factor in soil-adjusted vegetation index (savi) for aboveground living biomass estimation in arid grasslands. *Remote Sens. Environ.* **2018**, *209*, 439–445. [CrossRef]
41. Jordan, C.F. Derivation of leaf-area index from quality of light on the forest floor. *Ecology* **1969**, *50*, 663–666. [CrossRef]
42. Huete, A.R. A soil-adjusted vegetation index (savi). *Remote Sens. Environ.* **1988**, *25*, 295–309. [CrossRef]
43. Fu, G.; Shen, Z.X. Environmental humidity regulates effects of experimental warming on vegetation index and biomass production in an alpine meadow of the northern tibet. *PLoS ONE* **2016**, *11*, e0165643. [CrossRef]
44. Nembrini, S.; König, I.R.; Wright, M.N. The revival of the gini importance? *Bioinformatics* **2018**, *34*, 3711–3718. [CrossRef] [PubMed]
45. Boulesteix, A.-L.; Bender, A.; Bermejo, J.L.; Strobl, C. Random forest gini importance favours snps with large minor allele frequency: Impact, sources and recommendations. *Briefings Bioinform.* **2012**, *13*, 292–304. [CrossRef] [PubMed]
46. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef] [PubMed]
47. Huang, N.; Lu, G.; Xu, D. A permutation importance-based feature selection method for short-term electricity load forecasting using random forest. *Energies* **2016**, *9*, 767. [CrossRef]
48. Quinlan, R. *C4.5: Programs for Machine Learning Morgan Kaufmann*; Kluwer Academic Publishers: San Francisco, CA, USA, 1993.
49. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
50. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
51. Haq, M.A. Smotednn: A novel model for air pollution forecasting and aqi classification. *Comput. Mater. Contin.* **2022**, *71*, 1.
52. Haq, M.A. Cdlstm: A novel model for climate change forecasting. *Comput. Mater. Contin.* **2022**, *71*, 2363–2381.
53. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
54. Rossel, R.V.; Behrens, T.; Ben-Dor, E.; Brown, D.; Demattê, J.; Shepherd, K.D.; Shi, Z.; Stenberg, B.; Stevens, A.; Adamchuk, V.; et al. A global spectral library to characterize the world's soil. *Earth-Sci. Rev.* **2016**, *155*, 198–230. [CrossRef]
55. Rossel, R.V.; McGlynn, R.; McBratney, A. Determining the composition of mineral-organic mixes using uv–vis–nir diffuse reflectance spectroscopy. *Geoderma* **2006**, *137*, 70–82. [CrossRef]
56. Alley, W.M. The palmer drought severity index: Limitations and assumptions. *J. Appl. Meteorol. Climatol.* **1984**, *23*, 1100–1109. [CrossRef]
57. Zeng, N.; Ren, X.; He, H.; Zhang, L.; Li, P.; Niu, Z. Estimating the grassland aboveground biomass in the three-river headwater region of china using machine learning and bayesian model averaging. *Environ. Res. Lett.* **2021**, *16*, 114020. [CrossRef]
58. You, Y.; Wang, S.; Ma, Y.; Wang, X.; Liu, W. Improved modeling of gross primary productivity of alpine grasslands on the tibetan plateau using the biome-bgc model. *Remote Sens.* **2019**, *11*, 1287. [CrossRef]