Open Access

Advances in Environmental and Engineering Research

LIDSEN Publishing Inc.

Original Research

Utilizing Machine Learning to Map Aquatic Weed Yields for Biochar Production: A Case Study in the Bangweulu Wetlands, Zambia

Fabian Banda [1], Leonard Simukoko [1], Mulenga Kalumba [2, *], Mwansa Kaoma [2]

1. The University of Zambia, Technology Development and Advisory Unit, P.O. BOX 32379, Lusaka, Zambia; E-Mails: fabian.banda@unza.zm; leonard.simukoko@unza.zm
2. The University of Zambia, School of Engineering, Department of Agricultural Engineering, P.O. BOX 32379, Lusaka, Zambia; E-Mails: mulenga.kalumba@unza.zm; mwansa.kaoma@unza.zm

* **Correspondence:** Mulenga Kalumba; E-Mail: mulenga.kalumba@unza.zm

Academic Editor: Islam Md Rizwanul Fattah

Special Issue: Sustainable Biofuel & Bioenergy Production from Biomass & Biowaste Feedstocks

Adv Environ Eng Res
2025, volume 6, issue 2
doi:10.21926/aeer.2502020

Received: September 26, 2024
Accepted: April 14, 2025
Published: April 17, 2025

Abstract

Aquatic weeds present significant ecological and socio-economic challenges in the Bangweulu Wetlands of northern Zambia, where their proliferation disrupts aquatic ecosystems, impedes fishing activities, and affects local livelihoods. Despite these challenges, aquatic weeds also offer a unique opportunity for sustainable biochar production, a clean alternative cooking fuel that can alleviate pressure on diminishing forest resources. This study explores the application of machine learning (ML) techniques to estimate and map the spatial distribution of aquatic weed biomass, thereby enabling more efficient and strategic harvesting for biochar production. The research objectives included field-based measurement of aquatic weed biomass, analysis of environmental covariates, evaluation of four machine learning models for yield prediction, and the generation of spatial yield distribution maps. Among the tested models, Gradient Boosted Regression Trees (BRT) demonstrated superior performance, achieving an R^2 of 0.63, a Mean Absolute Error (MAE) of 0.08, and a Root Mean Square Error (RMSE) of 0.29. Key predictive variables included remote sensing-derived vegetation indices (LAI, EVI, NDVI), climate parameters, and topographic derivatives from Digital Elevation

Models (DEMs). Seasonal biomass yield predictions ranged from 0.70 kg to 1.18 kg per square meter, highlighting significant spatial and temporal variability. The ML-driven yield maps enable precision harvesting, which can enhance operational efficiency, reduce labor and fuel costs, and minimize environmental disturbance. Moreover, by facilitating the conversion of invasive biomass into biochar, the approach contributes to circular economy principles, reduces greenhouse gas emissions associated with traditional biomass use, and supports energy access in underserved rural areas. Overall, the integration of ML-based yield estimation into biochar production planning represents a scalable and data-driven solution that bridges environmental restoration with sustainable energy generation. The study's methodology and findings offer valuable insights for policymakers, conservationists, and clean energy advocates aiming to harness natural resources more responsibly.

Keywords
Aquatic weed yield mapping; biochar production for cooking fuel; environmental covariates; machine learning models; spatial quantification

1. Introduction

The Bangweulu Wetlands, situated in northern Zambia, comprise five islands with a population exceeding 20,000 people. Residents on these islands travel more than 40 km on waterways from the mainland (Samfya district), where they get their essential commodities. One of the most important commodities used to meet their cooking energy demands is traditional wood fuel, in the form of firewood (used in three-stone stoves) and charcoal (used in inefficient metal braziers). Because these islands lack forests to provide a supply of woodfuel for firewood and charcoal production, obtaining these woodfuel sources is costly. The cost of acquiring wood fuel is higher on these islands compared to rural mainland Zambia. Despite the recognized environmental, socioeconomic, health, and gender-related disadvantages associated with woodfuel use for cooking [1-3], residents of the islands currently have no viable access to alternative cooking fuels. It is widely reported that the use of woodfuel, which is burned in open fires or rudimentary cookstoves (as is the case in Bangweulu wetlands), leads to the release of harmful pollutants and greenhouse gases [4-7]. The repercussions of these practices go beyond health risks and climate change; they also have a significant influence on natural ecosystems. Thus, if there is any hope of meeting the country's climate, biodiversity, and sustainable development targets, there is a need to urgently expand access to clean cooking solutions in these areas.

Given the above, efforts toward sustainable development in the Bangweulu Wetlands will continue to be jeopardized until access to modern cooking solutions is improved. In this case, as adopted from the IEA [8], access to clean cooking solutions means access to (and primary use of) modern fuels and technologies, including natural gas, liquefied petroleum gas (LPG), electricity, and biogas, or improved biomass cookstoves (ICS), that have considerably lower emissions and higher efficiencies than traditional three-stone fires for cooking. Using the Multi-Tier Framework (MTF) approach, Luzi et al. [9] categorized these cooking solutions for Zambia into Tiers. The approach measures access to modern energy cooking solutions as a spectrum ranging from Tier 0 (no access)

to Tier 5 (full access) through six attributes: cooking exposure, cooking efficiency, convenience, availability of fuel, affordability, and safety of the primary cookstove. It captures the multidimensional nature of energy access and the vast range of technologies and sources that can provide energy access while accounting for the wide differences in user experience. Tier 0 and Tier 1 apply to firewood used in a three-stone stove and charcoal used in traditional braziers, respectively. Biomass-based fuels utilized in ICS, on the other hand, are classified as Tier 2. Tiers 3 and 4 include LPG, natural gas, and biogas used in gas stoves. Finally, electric stoves are classified as Tier 5.

Nearly all households in the rural areas of Zambia, including those in Bangweulu Wetlands, are mainly in Tier 0 or with very few in Tier 1. They are primarily constrained in these tiers by Cooking Exposure - one of the MFT attributes that assesses personal exposure to pollutants from cooking activities. Therefore, the ideal solution would be to promote clean fuel stoves in these areas, such as electric or LPG stoves, which emit little to no household pollutant emissions [10-12]. However, their promotion, especially in these rural areas, is hindered by two other MFT attributes - Affordability and Fuel Availability. Currently, apart from extremely limited access to electricity and LPG, respectively, these cookstove-fuel systems are unaffordable to residents of these areas (and most rural areas in Zambia) who have very low purchasing power [13, 14].

Hence, Luzi et al. [9] recommended that, while promoting these clean-fuel stoves in rural Zambia should be the long-term goal, a short-term solution would be to increase the adoption of improved biomass cookstoves as primary cooking solutions. In comparison to Tier 0 and 1 cookstove-fuel systems, they generate much fewer household pollutants and are more convenient to use [6, 12]; and are more affordable than the cookstove-fuel systems in Tiers 3 to 5. In addition, the use of ICS results in minimal disruption in cooking practices, and households can rely on existing fuel (wood or charcoal). Thus, adoption rates can increase faster than for clean-fuel stoves [9]. However, in the Bangweulu Wetlands, their promotion may be hindered by a lack of conventional sustainable biomass, primarily agricultural and forest residues. This is mainly attributed to the limited agriculture and forest industries that generate these residues in these areas.

Fortunately, aquatic weed is a readily available feedstock in these areas that can be exploited for the production of biochar briquettes, an alternative cooking fuel that may be used in ICS. According to recent research, aquatic weed, which comprises the main forms of energy molecules, including cellulose, hemicellulose, and lignin, is an excellent feedstock for biochar production (an immediate energy carrier for biochar briquettes) [15-19]. These studies report that about 50 to 73% of biochar can be produced from aquatic weeds through three main thermochemical processes - pyrolysis, gasification, or hydrothermal carbonization. There are notable benefits of using weed for bioenergy (in this biochar) production in the Bangweulu Wetlands. First, because of its effective utilization of nutrients and solar energy, this plant has a high rate of proliferation and hence produces enormous volumes of biomass [16], a much-needed feedstock for biochar production. Second, aquatic weeds do not compete directly with agricultural land resources, which are scarce in the Bangweulu Wetlands. Hence, its availability for biochar production does not interfere with agricultural production. Finally, by using aquatic weeds as biochar feedstock, the enormous capital and human resources required to safely manage and dispose of aquatic weeds in the wetlands can be greatly reduced.

While the advantages of using aquatic weeds in the Bangweulu Wetlands for biochar production are clear, a critical question arises: What is the actual quantity of these weeds available in the region?

Machine learning (ML) models provide a promising method to estimate the abundance of aquatic weeds, enabling accurate assessment of their availability for biochar production. Although many countries, such as Zambia, have no means of mapping aquatic weed yields in Wetland areas, moreover, very few researchers have used advanced technologies such as ML for spatial mapping purposes in Zambia [20, 21]. While previous studies have explored biomass and biochar production using ML, the novelty of this study lies in its context-specific application of ML to the Bangweulu Wetlands—a region where such data-driven environmental-energy integration is virtually non-existent. In Zambia, there has been minimal to no use of ML for mapping aquatic weed yield, and even fewer studies linking this information to biochar production as an alternative clean fuel source. Thus, this research introduces a first-of-its-kind approach in the Zambian wetland context, where ML is applied to spatially estimate and map aquatic weed biomass for targeted harvesting and energy planning. Furthermore, the study contributes a unique methodological innovation by integrating remote sensing indices (MODIS EVI, LAI, NDVI), climate data, and topographical variables into advanced ML models—Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Gradient Boosted Regression Trees (BRT)—to predict aquatic weed yield across space and time. This approach is not only novel for Zambia but also showcases how state-of-the-art AI/ML technologies can be operationalized for clean energy access in remote, data-scarce environments. By developing spatial yield maps, this research enables precision harvesting of aquatic weeds, which enhances operational efficiency, reduces harvesting costs, and minimizes ecosystem disruption. The result is a scalable, sustainable, and context-appropriate clean cooking solution that bridges environmental management with energy access. It fills a crucial knowledge and technology gap, offering an evidence-based pathway toward both climate adaptation and rural energy security.

Some examples of ML models include the Random Forest (RF), which is an ensemble learning method that combines multiple decision trees to make predictions and can also handle a large number of input variables and capture complex relationships between the input features and weed yield. It has been widely used in remote sensing and ecological studies to predict vegetation characteristics, including weed biomass [22-25]. For instance, Li et al. [26] & Zhu et al. [27] developed an RF regression-based biochar yield and carbon content prediction model using 245 datasets that covered various biomass feedstocks and process operating conditions. Their work achieved a coefficient of determination (R^2) value of 0.86 and 0.85 for predicting biochar yield and carbon content, respectively. Another type of ML model that can be used for quantifying aquatic weed yields is the Support Vector Machine (SVM). It is a supervised learning algorithm that can be used for classification or regression tasks. SVM has been applied to predict weed biomass based on remote sensing data and environmental variables. It works by finding an optimal hyperplane that maximally separates the data points corresponding to different weed biomass classes or regression values [28-32]. Bakhshipour & Jafari [33], Pereira et al. [34], and Tao & Wei [35] explored the accuracy of the SVM model and showed that the overall classification accuracy of Artificial Neural Network (ANN) was 92.92%, where 92.50% of weeds were correctly classified. Higher accuracies were obtained when the SVM was used as the classifier, with an overall accuracy of 95.00%, whereas 93.33% of weeds were correctly classified. Artificial Neural Networks (ANN) can handle sequential data, and they have been applied in time-series analysis of aquatic weed growth, considering factors such as historical weed biomass, environmental conditions, and seasonal patterns. ANN can capture temporal dependencies and make predictions based on the previous states of the sequence [34, 36-

38]. Some recent work achieved an improved R^2 = 0.92 for predicting biochar yield using an ANN model [39]. Finally, Boosted Regression Trees (BRT) is an ensemble learning technique that iteratively builds a strong predictive model by combining multiple weak models, such as decision trees. Boosting algorithms such as BRT have been used for weed biomass prediction, leveraging their ability to handle complex interactions between features and capture non-linear relationships [40-43]. Another work explored the accuracy of the BRT ML model for predicting biochar yield and found an R^2 value of 0.84 based on 91 training datasets [44].

Technologies such as ML and artificial intelligence (AI), deep learning, cloud computing, the use of drones, and edge computing can be used to get information and process it, and help raise productivity, improve quality, and ultimately increase awareness for spatially quantifying aquatic weed yield for the production of Biochar as an alternative cooking fuel in the Bangweulu Wetlands [21, 45]. By using ML and AI technology, we can now have better results to know what is going to happen, and where in the Wetlands, while promoting a data-driven approach for spatially quantifying aquatic weeds. Input features/independent variables or environmental covariates for the ML models may include spatial data layer maps such as remote sensing imagery e.g. the MODIS Enhanced Vegetation Index (EVI) [46], the MODIS Leaf Area Index (LAI) [46, 47], and the MODIS Normalized Vegetation Index (NDVI) [46], climate variables e.g. Rainfall, Potential Evapotranspiration (PET), and Surface Temperature [48], and a Digital Elevation Model (DEM) and its derivatives could be processed to show exactly when, where, and establish the spatial variation of aquatic weeds in the Bangweulu Wetlands. The study aimed to develop a methodology that utilizes machine learning techniques to accurately estimate and map the yield of aquatic weeds across the wetlands. This approach will enable efficient and targeted harvesting of the weeds for biochar production, promoting the use of biochar as an alternative cooking fuel and addressing the weed problem sustainably. This research aims to optimize the utilization of aquatic weeds in the Bangweulu Wetlands for biochar production as an alternative cooking fuel. To achieve this, the specific objectives are: (i) **Design and Implementation:** Conduct a systematic sampling campaign in selected locations within the Bangweulu Wetlands to collect aquatic weed samples and measure the biomass yield on a dry basis through laboratory analysis, (ii) **Relationship Investigation:** Analyze the relationship between the dry biomass yield of aquatic weeds and various environmental covariates, including remote sensing indices (MODIS EVI, LAI, and NDVI), climate variables (rainfall, potential evapotranspiration, and surface temperature), and topographical factors derived from the Digital Elevation Model (DEM), (iii) **Model Evaluation:** Compare the performance of four machine learning models—Artificial Neural Network (ANN), Gradient Boosted Regression Trees (BRT), Random Forest (RF), and Support Vector Machine (SVM)—in predicting the dry biomass yield of aquatic weeds based on the environmental covariates identified, and (iv) **Spatial Estimation and Mapping:** Utilize the best-performing machine learning model to estimate the total dry biomass yield of aquatic weeds across the entire Bangweulu Wetlands. Develop detailed digital yield maps representing monthly and seasonal variations across the wetlands.

2. Methodology

2.1 Study Area

The Bangweulu Wetlands, located in Zambia, are among the most fascinating ecological study areas in Africa. It is one of the largest wetland ecosystems in Africa and a significant ecological

hotspot. Researchers and scientists have long been intrigued by the unique biodiversity and ecological processes that occur within this wetland system. Aquatic weeds pose a significant challenge in the Bangweulu Wetlands, affecting the ecological balance and impeding local communities' livelihoods [49]. To develop a predictive model that can accurately estimate the biomass of aquatic weeds and identify suitable areas for sustainable harvest, three items are required, i) the dependent variable, which is the amount of aquatic biomass weed yield on a dry basis, ii) the input features/independent variables which were the environmental covariates, and iii) the best-performing ML model.

2.2 Design and Implementation of an Aquatic Weeds Sampling Campaign Across the Bangweulu Wetlands

To develop a robust dataset for training and evaluating the performance of machine learning models, aquatic weed biomass data (dry basis) were collected from 154 georeferenced sampling points distributed across the Bangweulu Wetlands. The selection of this sample size was informed by previous ecological field studies conducted in similar large wetland ecosystems and the spatial heterogeneity of aquatic vegetation in the Bangweulu Wetlands, as well as practical field constraints such as accessibility and resource availability. Although a formal power analysis was not conducted, the sample size was deemed sufficient to capture the spatial variability of aquatic weed biomass across the study area while ensuring statistical reliability for model training and validation. A random transect-based sampling design was adopted to ensure unbiased and representative coverage of the diverse aquatic habitats within the Wetlands. In December 2022, a series of randomly oriented transects were established across different zones of the wetland system. Along each transect, sample plots were randomly located at predetermined intervals to avoid clustering and reduce spatial autocorrelation. At each of these locations, aquatic weed biomass was harvested using a 0.04 m^2 quadrat. The geographic coordinates of each sampling point were recorded with high precision using a Differential Germin GPS device, ensuring spatial accuracy for later integration with remote sensing and environmental covariate data (Figure 1). Following field collection, samples were transported to the laboratory where they were carefully sorted to remove debris, non-target vegetation, and other contaminants. The cleaned samples were rinsed with water to eliminate sediment, then oven-dried at controlled temperatures ranging from 50°C to 70°C until a consistent dry weight was achieved. The dry biomass yield (n = 154) was then measured for each sample and compiled into a dataset used in the subsequent machine learning analysis. This sampling strategy allowed for the generation of a high-quality, spatially explicit dataset that supports accurate modeling of aquatic weed yield across the Bangweulu Wetlands.

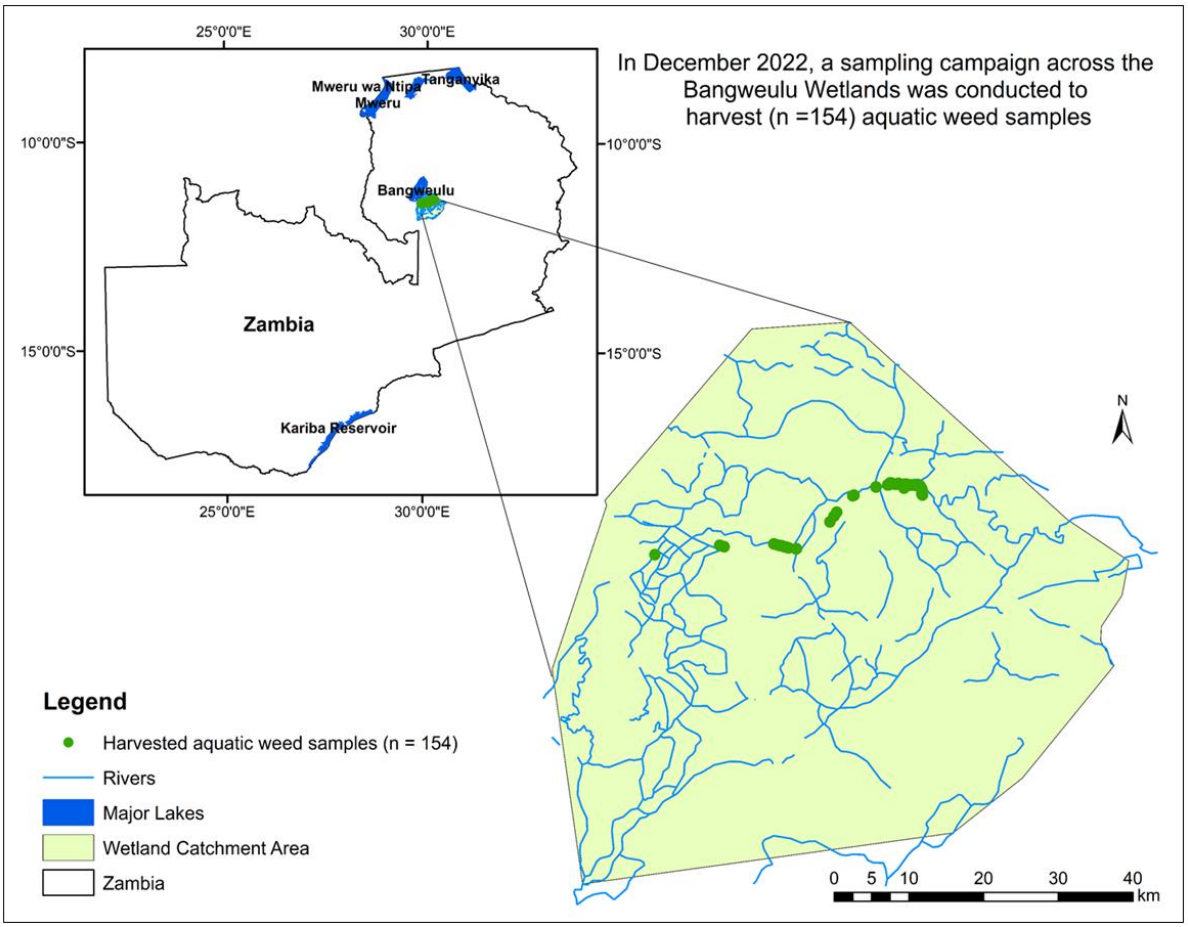

Figure 1 The Bangweulu Wetlands in Zambia and the location of 154 points where aquatic weeds were harvested, and for which the amount of aquatic biomass weed yield on a dry basis for (n = 154) was measured in the Laboratory.

2.3 Investigation of the Relationship between the Amount of Aquatic Biomass Weed Yield on a Dry Basis and the Environmental Covariates

In spatial modeling studies, it is important to have a good number of environmental covariates/independent variables or input features for optimal models to be developed; hence, in this study, 43 input features or environmental covariates/independent variables were identified from various sources listed in Table 1. According to [20, 50, 51], in spatial modeling studies, the use of many environmental covariates is highly encouraged. These variables are available as spatial coverages, each with its spatial resolution ranging from 10 to 1000 m, over the whole Bangweulu Wetland Catchment area. All of the covariate layers were projected in the same cartographic reference system (WGS84/UTM zone 35S) and resampled to a 30 m spatial resolution using the bilinear approach for continuous covariates [52]. A selection of the environmental covariates (six variables) used is displayed in Figure 2 and Table 1.

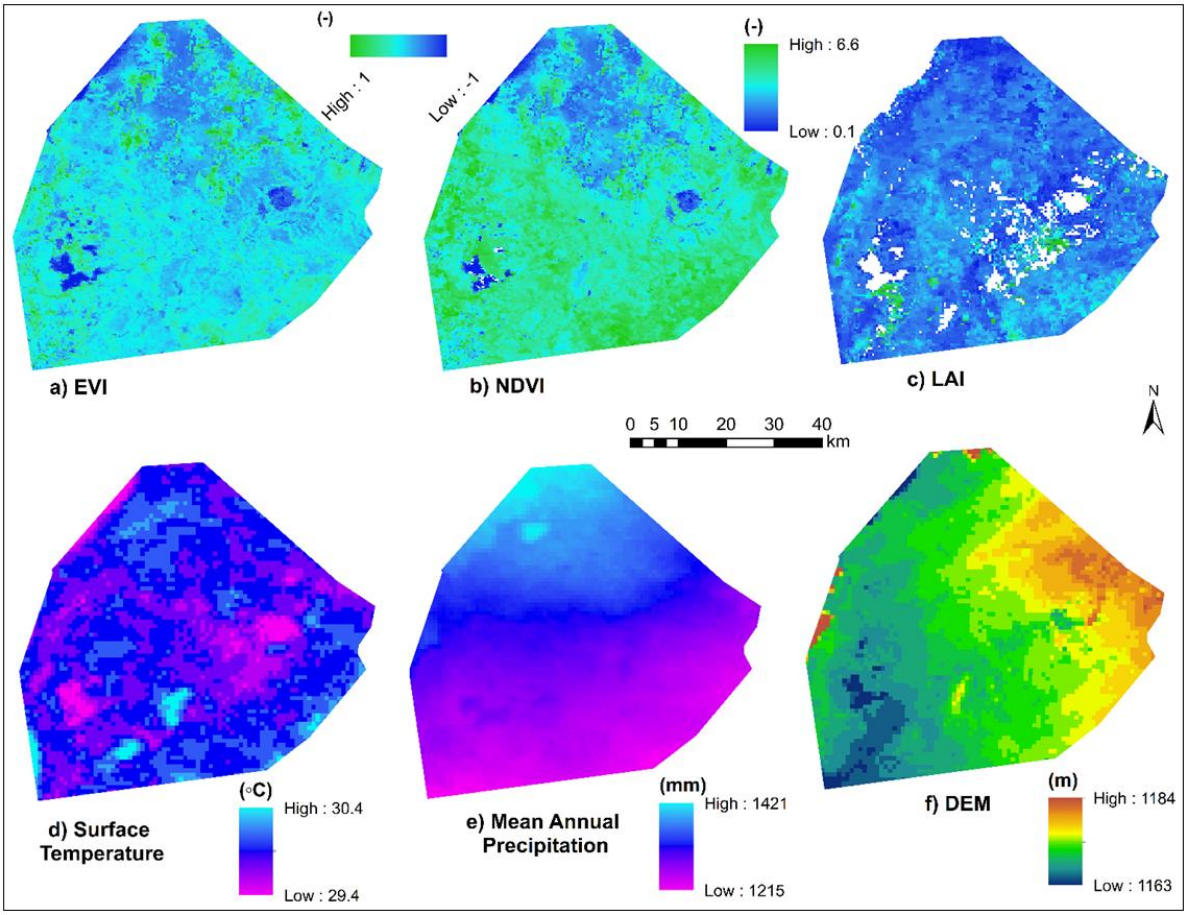

Figure 2 Six out of the total 43 identified and examined environmental covariates: a) Enhanced Vegetation Index (EVI), b) Normalized Vegetation Index (NDVI), c) Leaf Area Index (LAI), d) Surface Temperature, e) Mean Annual Precipitation, and f) the Digital Elevation Model (DEM).

Table 1 Environmental covariates and their abbreviations, type, spatial resolution, units, and range of values.

Abbreviation	Covariate	Type of data	Spatial resolution	Units	Range of values	Source
Climate factor						
PET	Potential Evapotranspiration	Continuous	"	"	1000-2500	Global Aridity Index and PET Database [53, 54]
BIO1	Annual Mean Temperature	"	"	°C	10-26	WorldClim database [48]
BIO2	Mean Diurnal Range	"	"	"	7-22	"
BIO3	Isothermality	"	"	%	52-74	"
BIO4	Temperature Seasonality	"	"	-	146-386	"
BIO5	Max Temperature of Warmest Month	"	"	°C	17-37	"
BIO6	Minimum Temperature of Coldest Month	"	"	"	2-17	"
BIO7	Temperature Annual Range	"	"	"	14-30	"
BIO8	Mean Temperature of Wettest Quarter	"	"	"	12-29	"
BIO9	Mean Temperature of Driest Quarter	"	"	"	8-24	"
BIO10	Mean Temperature of Warmest Quarter	"	"	"	12-30	"
BIO11	Mean Temperature of Coldest Quarter	"	"	"	7-23	"
BIO12	Annual precipitation	"	1000 m	mm	400-2200	"
BIO13	Precipitation of the Wettest Month	"	"	mm	150-650	"
BIO14	Precipitation of the Driest Month	"	"	"	0-40	"
BIO15	Precipitation Seasonality	"	"	"	75-136	"
BIO16	Precipitation of the Wettest Quarter	"	"	"	200-1340	"
BIO17	Precipitation of the Driest Quarter	"	"	"	0-126	"
BIO18	Precipitation of the Driest Quarter	"	"	"	74-760	"
BIO19	Precipitation of the Coldest Quarter	"	"	"	0-160	"

M01MOD4	Mean Monthly Surface Reflectance	"	250 m	-	0-1	MODIS Surface Reflectance [55]
M06MOD4	"	"	"	"	"	"
M12MOD4	"	"	"	"	"	"
X101MOD4	"	"	"	"	"	"
X106MOD4	"	"	"	"	"	"
X112MOD4	"	"	"	"	"	"
T01MOD	Mean Monthly Surface Temperature	"	"	℃	29-31	MODIS Surface Temperature [55]
T06MOD3	"	"	"	"	"	"
T12MOD3	"	"	"	"	"	"

2.4 Machine Learning (ML) Models

In this study, four machine learning (ML) models—Artificial Neural Networks (ANN), Gradient Boosted Regression Trees (BRT), Random Forest (RF), and Support Vector Machines (SVM)—were selected based on their suitability for handling complex, high-dimensional data and their proven efficacy in environmental and biomass prediction studies [56, 57]. Each model offers distinct advantages:

Gradient Boosted Regression Trees (BRT): BRT models were chosen for their ability to handle non-linear relationships and interactions among predictors, which are common in environmental datasets. Their iterative approach allows for the progressive improvement of predictions by minimizing error, making them highly effective for modeling spatial variability. However, BRT models are computationally intensive and require careful parameter tuning to avoid overfitting [58].

Random Forest (RF): RF models are robust ensemble learners that combine multiple decision trees to improve prediction accuracy. They excel in handling large numbers of input variables and can measure variable importance, providing insights into key environmental covariates influencing biomass yield. Despite their robustness, RF models can sometimes oversimplify relationships, which might limit their predictive power in highly non-linear scenarios [59, 60].

Support Vector Machines (SVM): SVM models were selected for their ability to model complex data with high-dimensional features. They are particularly useful in cases where the relationship between predictors and the response variable is not explicitly linear. However, their performance can be sensitive to the choice of kernel functions and hyperparameters, requiring extensive tuning [56].

Artificial Neural Networks (ANN): ANN models mimic the human brain's neural structure and are highly adaptable to capturing intricate patterns in data. Their ability to handle temporal and spatial dependencies makes them valuable for biomass prediction. However, their performance heavily depends on the quality of training data, and they are often referred to as "black box" models due to their lack of interpretability [57]. The selection of these models was guided by their complementary strengths, ensuring that the study leveraged diverse approaches to achieve accurate predictions of biomass yield. By comparing their performance using metrics such as the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE).

2.5 Performance Evaluation of Estimated Aquatic Biomass Weed Yield on a Dry Basis Using Four ML Models

R programming software, version 4.3.1 [61], was used for all the data preprocessing, ML model training, testing, and validation. All 43 environmental covariates were transformed to the *GeoTiff* raster format. The 43 covariate layers were first reprojected into the same WGS84/UTM zone 35S coordinate system and then resampled to a 30 m spatial resolution. In R, using the *raster package* and *stack function*, a *RasterStack* dataset consisting of all 43 environmental covariates listed in Table 1 was created. By overlaying the *RasterStack dataset*, which contains all the 43 environmental covariates, with the georeferenced point dataset containing the measured values of the response variable—that is, the aquatic biomass weed yield on a dry basis (n = 154)—shown in Figure 1, a *RegressionMatrix* was produced using the R software's *extract function*. As a result, the values of a response variable (n = 154) and the associated pixel values of each of the 43 environmental variables that were employed in this study made up this *RegressionMatrix*. The parsimonious Multiple Linear

Regression (MLR) model was proposed through a systematic and iterative methodology implemented in R. The process began with the construction of a RegressionMatrix, which combined the response variable (aquatic biomass weed yield on a dry basis) measured at 154 georeferenced points across the Bangweulu Wetlands with 43 identified environmental covariates. These covariates were preprocessed to ensure uniform spatial resolution and coordinate reference systems. Using this dataset, the backward-stepwise selection method based on the Akaike Information Criterion (AIC) R package [62] was applied. This approach is a recognized statistical technique for model optimization, aiming to balance model complexity and goodness of fit by penalizing additional covariates. Specifically, the process involved evaluating 43 potential models iteratively, removing one covariate at a time, and calculating the AIC after each step. The iterative reduction of covariates continued until further exclusion no longer reduced the AIC value. This point marked the identification of the most parsimonious model. The final model, described as (1), comprised 27 covariates. These were selected for their contribution to minimizing the AIC, indicating they provided the best explanatory power for aquatic weed yield while maintaining simplicity and interpretability. The resulting model ensured the exclusion of redundant or less significant covariates, thereby enhancing its parsimony and predictive robustness. This systematic application of the AIC-based backward-stepwise selection process ensured that the final MLR model retained only the most critical covariates for explaining the variability in aquatic weed yield. This methodological rigor aligns with established practices in regression modeling and model selection in environmental research. This parsimonious model is given as:

$$\begin{aligned}\textbf{Aquatic Biomass Weed Yield} = {} & \mathrm{X}_{101}\mathrm{MOD}_4 + \mathrm{X}_{106}\mathrm{MOD}_4 + \mathrm{X}_{112}\,\mathrm{MOD}_4 + \mathrm{BIO}_{10} + \mathrm{BIO}_{11} \\ & + \mathrm{BIO}_{12} + \mathrm{BIO}_{13} + \mathrm{BIO}_{15} + \mathrm{BIO}_{16} + \mathrm{BIO}_2 + \mathrm{BIO}_3 + \mathrm{BIO}_4 \\ & + \mathrm{BIO}_5 + \mathrm{BIO}_8 + \mathrm{BIO}_9 + \mathrm{CRU} + \mathrm{CRV} + \mathrm{DEM} + \mathrm{EVI}_1\_\mathrm{DEC} \\ & + \mathrm{LAI}_1\_\mathrm{DEC} + \mathrm{LAI}_2\_\mathrm{DEC} + \mathrm{LAI}_3\_\mathrm{DEC} + \mathrm{LAI}_4\_\mathrm{DEC} \\ & + \mathrm{M}_{06}\mathrm{MOD}_4 + \mathrm{M}_{12}\mathrm{MOD}_4 + \mathrm{NDVI}_1\_\mathrm{DEC} + \mathrm{T}_{06}\mathrm{MOD}_3 \end{aligned} \quad (1)$$

2.6 Mapping of the Amount of Aquatic Biomass Weed Yield on a Dry Basis for the Whole Bangweulu Wetlands Using the Best-Performing ML Model

The sensibility of the solutions determined by the tuning process in R was ensured through a systematic and validated approach. Using the parsimonious model with 27 potential covariates as the foundation, a three-fold cross-validation method was employed to train, test, and validate the selected machine learning (ML) models. This approach is widely recognized in ML for its robustness in preventing overfitting and ensuring that the models generalize well to unseen data. Four ML models—Artificial Neural Network (ANN), Gradient Boosted Regression Trees (BRT), Random Forest (RF), and Support Vector Machine (SVM)—were evaluated. The meta-parameters for each model were optimized using the *mlr* R package, which employs automated tuning to balance model complexity and performance. The tuning process specifically adjusted key meta-parameters to achieve the best predictive accuracy for: 1) Artificial Neural Network (ANN): Meta-parameters such as the number of *neurons (15)* and *hidden layers (2)* were tuned to improve the network's ability to capture complex, non-linear relationships in the data, 2) Gradient Boosted Regression Trees (BRT): Parameters including *n.tree (1000), shrinkage (0.1)*, and *interaction depth (10)* were optimized to enhance the model's ability to handle interactions and non-linear effects, 3) Random Forest (RF): The *mtry (4)* and *n.tree (1000)* parameters were fine-tuned to achieve a balance between

computational efficiency and predictive accuracy, and 4) Support Vector Machine (SVM): The *gamma (88.2) and cost function (0.794)* parameters were carefully adjusted to improve the model's ability to manage high-dimensional feature spaces and non-linear separations. The rationale behind these parameter choices is grounded in the literature and practical experience with these ML models. For example, the ANN structure was designed to handle the complexity of interactions among the 27 covariates, while the BRT and RF parameters were optimized to exploit their ensemble learning strengths. The SVM parameters were tailored to maximize the separation between data points in the feature space. To ensure the sensitivity of the tuning outcomes, the performance of the models was evaluated using three key metrics: the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). The model with the best combination of high R^2, low MAE, and low RMSE was identified as the most suitable for the task. This rigorous and systematic tuning approach demonstrates that the solutions are not only statistically valid but also grounded in a clear understanding of the underlying ML algorithms and their application to spatial modeling tasks. By tailoring the meta-parameters to the specific characteristics of the dataset and covariates, the process ensures that the selected models are both reliable and interpretable. The structure of the Artificial Neural Network (ANN) was chosen based on a systematic tuning process aimed at balancing model complexity with predictive accuracy. This was achieved using the mlr package in R, which allows automated tuning of meta-parameters to optimize model performance.

The ANN was designed with 15 neurons and 2 hidden layers, as shown in Figure 3. These structural elements were selected following an iterative process that tested various configurations to identify the most effective architecture for capturing the relationships within the dataset. The choice of 15 neurons and 2 hidden layers was informed by the complexity of the relationships among the 27 covariates included in the parsimonious model. The environmental covariates comprised a mix of remote sensing indices, climate variables, and topographic features, which are known to exhibit non-linear and interdependent interactions. The multi-layered ANN structure was tailored to account for these interactions by enhancing the network's capacity to model complex patterns in the data. The selected structure of the ANN aligns with established practices in environmental modeling, where neural networks are often configured to handle multi-dimensional datasets with complex dependencies. The iterative tuning process and the robust evaluation framework confirm that the chosen architecture is both appropriate and effective for the task of estimating aquatic weed yield in the Bangweulu Wetlands.

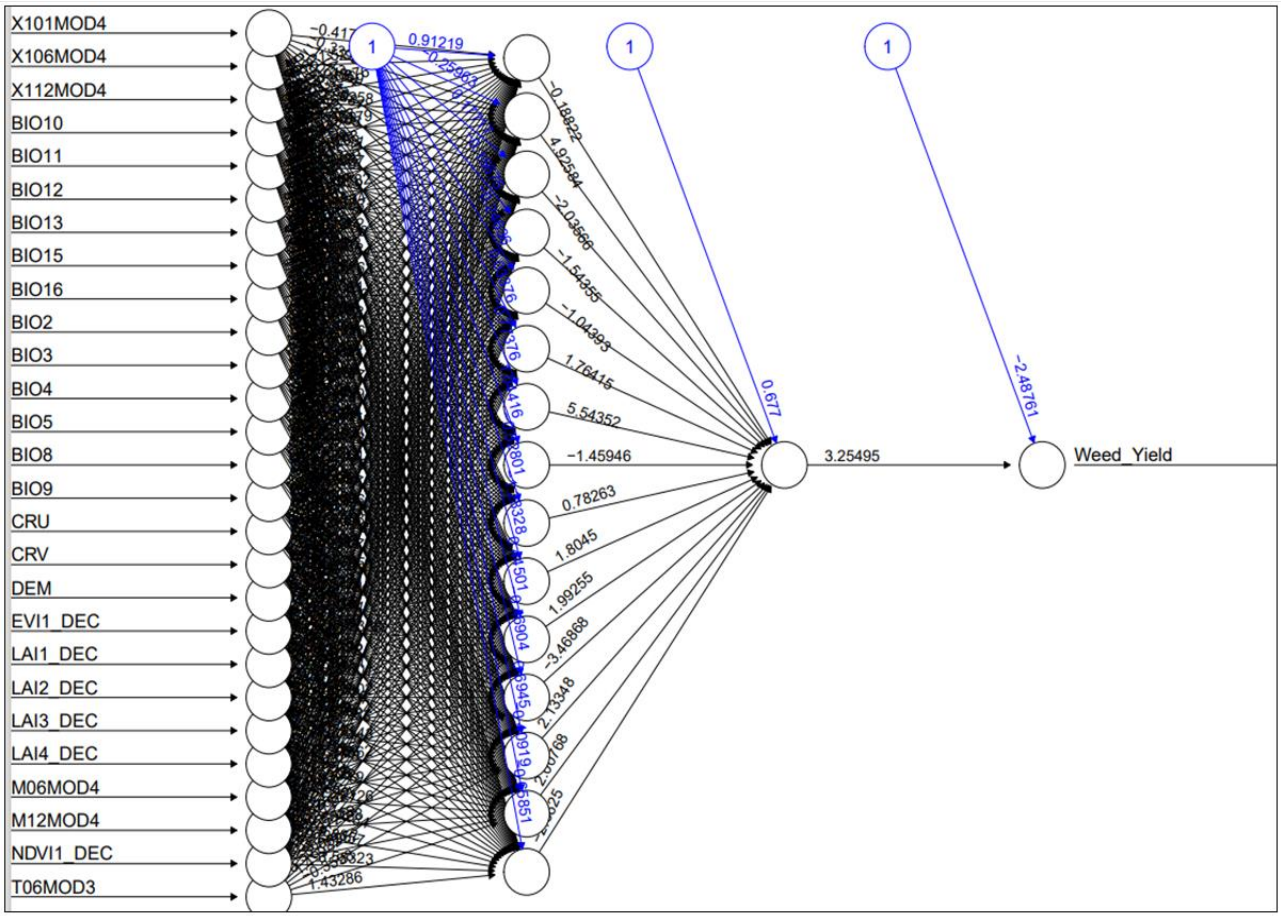

Figure 3 An architecture or structure of an ANN ML model with 15 neurons and 2 hidden layers for estimating or predicting aquatic biomass weed yield on a dry basis.

The best-performing model among the four ML models was subsequently used to estimate or predict the amount of aquatic biomass weed yield on a dry basis for the whole Bangweulu Wetlands catchment area for January, February up to December 2022. Furthermore, the 27 potential covariates that were selected in the model training and testing were ranked according to their order of importance relative to the particular response variable using the *relative importance function* of the random forest model.

3. Results and Discussion

The application of machine learning techniques for spatially quantifying aquatic weed yield offers several advantages. It provides a non-invasive and cost-effective approach to monitoring the wetlands' weed abundance, enabling timely and informed decision-making for biochar production. Additionally, the generated spatial maps can facilitate targeted interventions and resource allocation, enhancing the overall efficiency and sustainability of the biochar production process.

3.1 Ranking of Variables in Order of Importance

Figure 4 presents the results of a random forest model, displaying the 27 highest-ranked environmental covariates out of 43 candidate predictor variables listed in Table 1. The variables are ranked based on their importance, measured by the Mean Decrease Gini (Inc NodePurity) on the y-

axis. The IncNodePurity value indicates how much the model accuracy would decrease if a particular variable is left out. A higher mean decrease Gini score signifies a higher importance of the variable to the model.

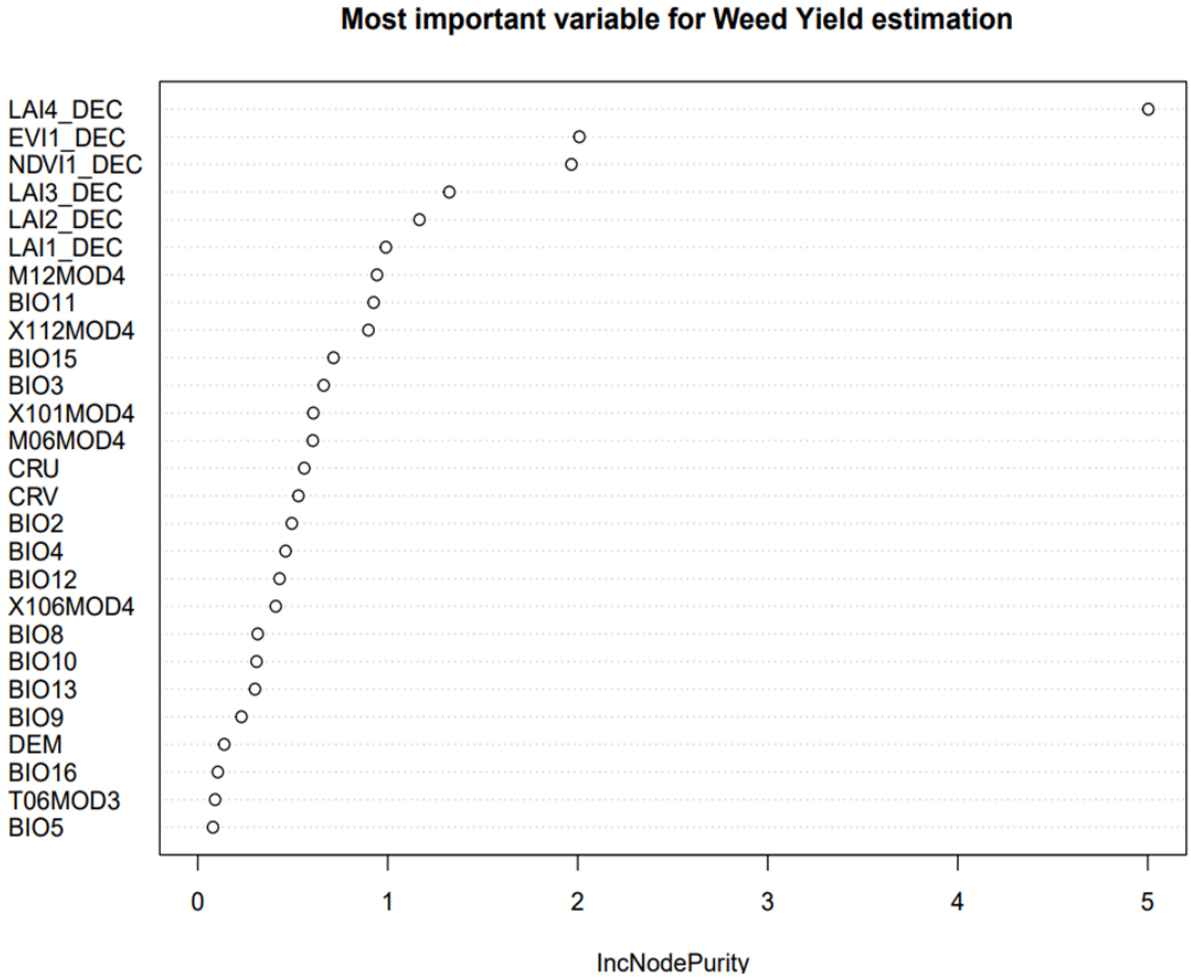

Figure 4 The 27 most important covariates for the aquatic weed yield response variable expressed in terms of Mean Decrease Gini (IncNodePurity), as derived from a random forest model. Abbreviations of the covariates are given in Table 1.

The results show that remote sensing indices, including the Leaf Area Index (LAI_4_DEC), Enhanced Vegetation Index (EVI_1_DEC), and Normalized Vegetation Index ($NDVI_1$_DEC), were identified as the most important predictors for estimating aquatic biomass weed yield on a dry basis. These indices provide information about the vegetation cover and health, which can be indicative of weed biomass. Climate variables such as Mean Monthly Surface Reflectance ($M_{12}MOD_4$), Mean Temperature of Coldest Quarter (BIO_{11}), Precipitation Seasonality (BIO_{15}), and Isothermality (BIO_3) were also found to be significant predictors. These climate variables capture information about the temperature and precipitation patterns, which can influence weed growth. Additionally, Digital Elevation Model (DEM) derivatives, specifically Local Upslope Curvature (CRU) and Downslope Curvature (CRV), were identified as important predictors, albeit not as influential as the remote sensing indices and climate variables. Overall, the findings align with the existing literature, which suggests that remote sensing indices and climate variables play crucial roles in estimating weed biomass [63-65]. The use of remote sensing indices, such as LAI, EVI, and NDVI, has been widely

documented in studies focusing on vegetation and biomass estimation [47, 66-68]. Similarly, climate variables have been recognized as important factors influencing weed growth and distribution [69-71]. The inclusion of DEM derivatives also complements previous studies that have highlighted the significance of topographic features in understanding wetland ecosystems [72-74].

3.2 Model Performance Evaluation

The performance evaluation of each model was assessed using three evaluation metrics: coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). Figure 5 provides a visual representation of the model performance based on these metrics using a three-fold cross-validation in R.

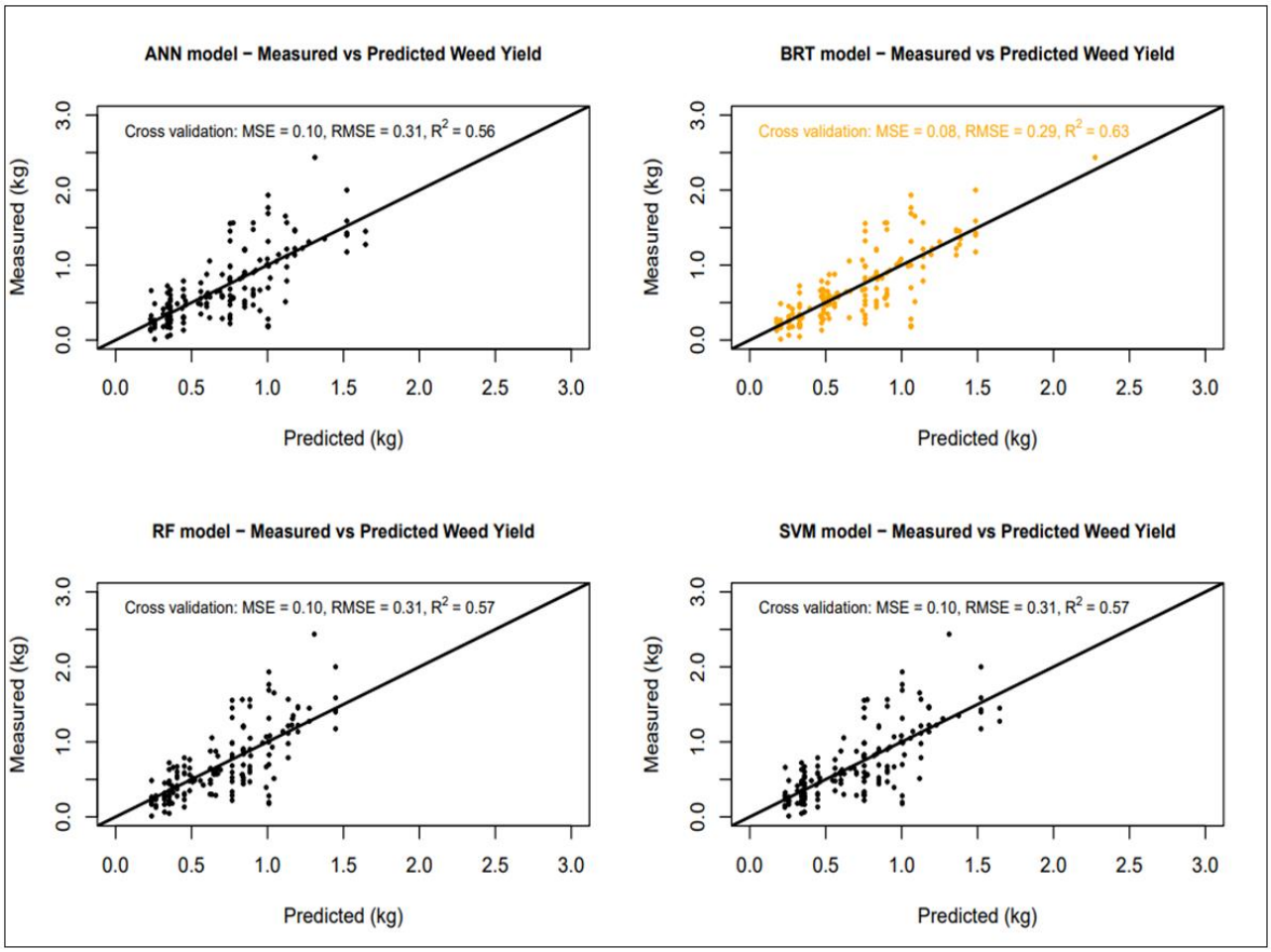

Figure 5 Coefficients of determination (R^2) for the four ML models applied to the three-fold cross-validation datasets in R. ANN, artificial neural network; BRT, gradient boosted regression trees; RF, random forest; SVM, support vector machine.

According to the results in Figure 5, the Gradient Boosted Regression Trees (BRT) model achieved the highest performance among the four ML models, with an R^2 value of 0.63, MAE of 0.08, and RMSE of 0.29. The Support Vector Machine (SVM) model followed closely behind with an R^2 of 0.57, MAE of 0.10, and RMSE of 0.31. The Random Forest (RF) model and Artificial Neural Network (ANN) model had similar performance, both achieving an R^2 of 0.57, MAE of 0.10, and RMSE of 0.31. In general, the performance of the ML models in this study appears to be consistent with previous

research. For example, Li et al. [7] & Zhu et al. [27] used 245 datasets covering different biomass feedstocks and process operating conditions to develop an RF regression-based biochar yield and carbon content prediction model. Their research produced R^2 for carbon content and biochar yield prediction values of 0.85 and 0.86, respectively. The BRT model's performance in this study, with an R^2 of 0.63, is relatively good. Based on 91 training datasets, a different study examined the accuracy of the BRT ML model in predicting biochar yield and found an R^2 value of 0.84 [44].

Research has shown that BRT models often exhibit strong predictive capabilities, particularly in situations where there are complex interactions among variables [75]. Therefore, in this study, the excellent performance of the BRT model in quantifying aquatic weed yield may be attributed to its ability to capture non-linear relationships and interactions within the dataset. The SVM model's performance, with an R^2 of 0.57, is also noteworthy. SVM is a widely used ML technique known for its ability to handle high-dimensional data and non-linear relationships. Previous works have shown SVM to be effective in various prediction tasks, including regression [75]. The results obtained in this study further support the efficacy of SVM for spatial quantification of aquatic weed yield. The RF model and ANN model achieved similar performance, both with an R^2 of 0.57. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions, while Artificial Neural Networks are computational models inspired by the human brain's neural structure. Both models have been widely applied in various domains, including agriculture and environmental sciences [76, 77]. The comparable performance of RF and ANN in this study suggests their suitability for predicting aquatic weed yield in the Bangweulu Wetlands. In summary, the results of this study demonstrate that the Gradient Boosted Regression Trees (BRT) model outperformed the other ML models, including Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN), in spatially quantifying aquatic weed yield. These findings align with the literature, where BRT models have shown strong predictive capabilities, while SVM, RF, and ANN models have also demonstrated effectiveness in similar prediction tasks. It is important to consider the specific context and dataset when comparing results, but overall, the performance of the ML models in this study is consistent with existing research. It is important to acknowledge that, while the ML models demonstrated satisfactory predictive performance, they are not without limitations. One common concern in ML applications is overfitting, where a model captures noise or anomalies in the training data rather than generalizable patterns, potentially reducing its performance on new, unseen data. Although cross-validation was employed to reduce this risk, the relatively modest R^2 values suggest that the models could still be sensitive to variations in the input data. Additionally, spatial autocorrelation effects—where nearby spatial observations are more similar than distant ones—can introduce bias into model estimates if not properly accounted for. This could result in inflated model performance metrics, especially in spatially clustered datasets like those used in this study. Incorporating spatial cross-validation or explicitly modeling spatial structure in future research would improve the robustness of predictive models in similar ecological contexts.

3.3 Aquatic Biomass Weed Yield Mapping

The study's findings included the generation of an aquatic biomass weed yield map (Figure 6), which visually represented the spatial distribution of aquatic weed yields across the Bangweulu Wetlands, and the presentation of the results as boxplots (Figure 7), illustrating the seasonal variations in the estimated yields. These results provided a comparative analysis between the

aquatic biomass weed yields estimated by the Gradient-Boosted Regression Trees (BRT) model and the measured yields reported during the study. The comparison demonstrated the alignment of the model's predictive estimates with the observed data, highlighting the BRT model's effectiveness in estimating aquatic weed yields in the study area. In the rainfall season, the estimated aquatic biomass weed yield varied from 0.01 to 1.50 kg with an average of 0.84 kg for November, 0.05 to 1.89 kg with an average of 1.13 kg for December, 0.13 to 1.79 kg with an average of 1.07 kg for January, 0.01 to 1.85 kg with an average of 1.14 kg for February, and 0.01 to 1.83 kg with an average of 1.10 kg for March. These estimates were compared to the measured aquatic biomass weed yield, which ranged from 0.01 to 2.40 kg with an average of 0.70 kg. In the dry season, the estimated aquatic biomass weed yield varied from 0.01 to 1.39 kg with an average of 0.80 kg for April, 0.01 to 1.46 kg with an average of 0.79 kg for May, 0.01 to 1.45 kg with an average of 0.79 kg for June, 0.01 to 2.03 kg with an average of 1.06 kg for July, 0.01 to 1.89 kg with an average of 1.15 kg for August, 0.01 to 1.91 kg with an average of 1.18 kg for September, and 0.01 to 1.57 kg with an average of 0.82 kg for October. These estimates were compared to the measured aquatic biomass weed yield, which ranged from 0.01 to 2.40 kg with an average of 0.70 kg.

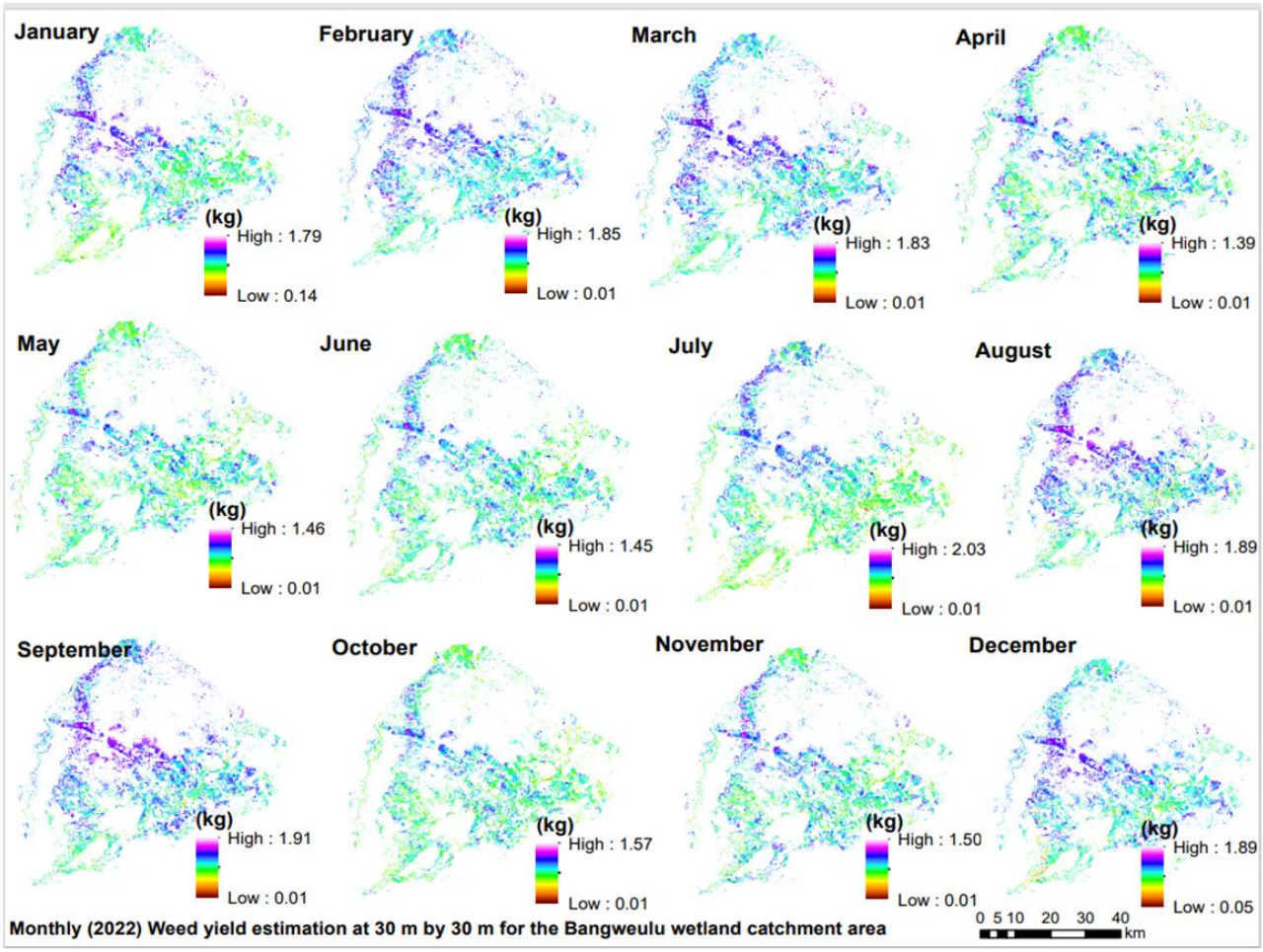

Figure 6 Aquatic Biomass Weed Yield maps for the year 2022 in the Bangweulu Wetlands. The data has a spatial resolution of 30 m, and the maps were generated using a Gradient-Boosted Regression Trees (BRT) model based on environmental covariates.

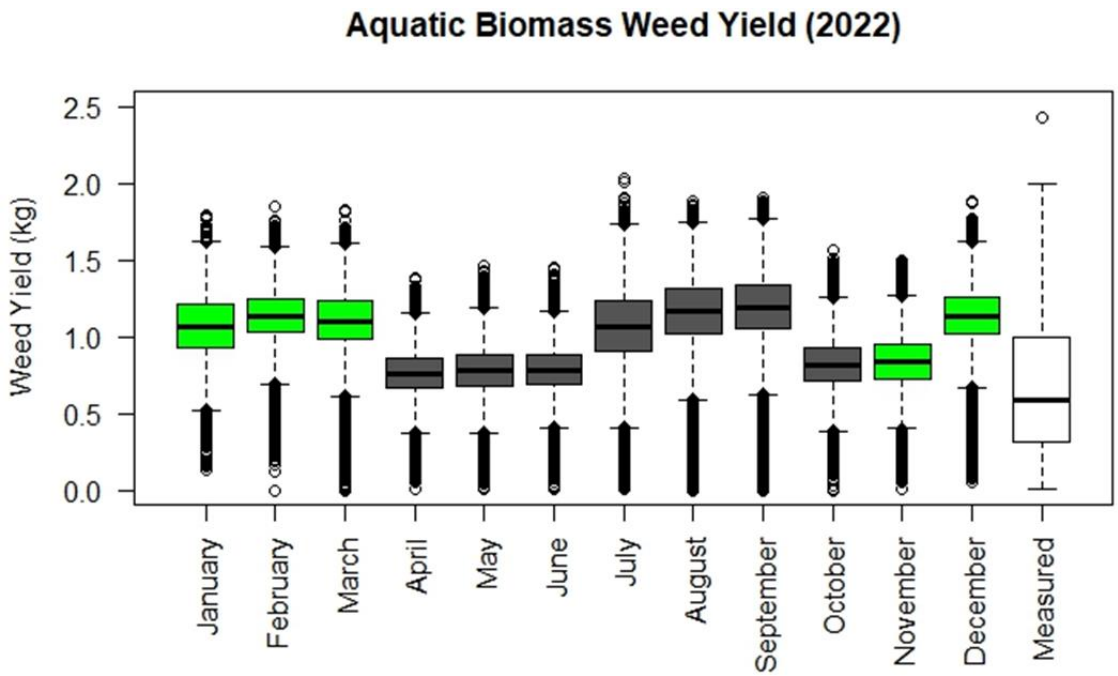

Figure 7 Aquatic Biomass Weed Yield maps for the year 2022 in the Bangweulu Wetlands. The data were generated using a Gradient-Boosted Regression Trees (BRT) model, whereby, the green boxplots are the Weed Yield (kg) estimates in the wet season, the grey boxplots are the Weed Yield (kg) estimates in the dry season, and the white boxplot is the measured Weed Yield (kg).

The results obtained through the machine learning approach, specifically the BRT model, enabled the estimation of aquatic biomass weed yield in the Bangweulu Wetlands. The developed yield map provided spatial information on the distribution of aquatic weed yield, while the box plots summarized the estimated and measured values for different months. Comparing the estimated aquatic biomass weed yield with the measured values, it is observed that the estimated yields generally show higher values compared to the measured yields. For both the rainfall and dry seasons, the average estimated yields were higher than the average measured yields in each respective month. However, without further information on the specific characteristics of the dataset, training methods, and validation techniques used in the study, it is challenging to make a direct comparison with what exists in the literature. The accuracy of the estimated yields and the performance of the BRT model cannot be fully assessed without considering the validation process and evaluating the model against independent datasets. To determine the best month to harvest aquatic weed yield in the Bangweulu Wetlands for biochar production, it is necessary to consider additional factors such as weed growth patterns, nutrient content at different times of the year, and logistical considerations. The study does not provide explicit recommendations regarding the best month for harvesting based on the obtained results. Further research and validation efforts are essential to validate the accuracy of the machine learning model and to compare the results with existing literature on aquatic weed yield in the Bangweulu Wetlands.

3.4 Comparative Analysis

The performance of the machine learning models in this study was benchmarked against findings from previous studies to evaluate consistency and highlight advancements. Table 2 presents a comparative summary of the model performances and key findings, including those reported in earlier works.

Table 2 Comparative Analysis of Model Performances and Key Findings.

Study/Model	Dataset Characteristics	Best Performing Model	R^2 Value	Key Findings
Current Study	154 georeferenced aquatic weed samples	BRT	0.63	BRT performed best due to its ability to model non-linear interactions among predictors.
Pathy et al. [44]	91 datasets on algal biochar yield	BRT	0.84	Highlighted BRT's capacity to handle complex biomass data.
Li et al. [7]	245 datasets on biochar yield	RF	0.85	RF excelled in predicting biochar yield and carbon content.
Zhu et al. [27]	Diverse biomass feedstock data	RF	0.86	RF demonstrated superior predictive accuracy for biomass yields across multiple conditions.
Herrero-Huerta et al. [41]	UAS-based multispectral data	BRT	0.78	Demonstrated BRT's robustness for vegetation biomass prediction.

This comparative analysis illustrates that the performance of the BRT model in this study, with an R^2 of 0.63, aligns with previous findings, although variations in R^2 values reflect differences in dataset size, type, and covariates used. Studies such as Pathy et al. [44] and Herrero-Huerta et al. [41] also emphasized the strength of BRT in capturing non-linear relationships and interactions. Meanwhile, studies by Li et al. [7] & Zhu et al. [27] highlighted Random Forest's suitability for biochar yield prediction, further supporting the importance of ensemble learning methods in biomass estimation. By presenting these comparisons, the study underscores the relevance and reliability of BRT for spatially quantifying aquatic weed yields, particularly in environments with complex predictor-response relationships, such as the Bangweulu Wetlands. This analysis not only validates the study's findings but also positions the developed methodology as a robust and transferable approach for similar ecosystems.

3.5 Sampling Campaign Done in a Different Month than December

If the sampling campaign had been conducted in a different month than December, the dataset generated for model training, testing, and validation would likely reflect seasonal variations in aquatic weed growth. December represents the peak of the rainy season in the Bangweulu Wetlands, a period associated with optimal growth conditions for aquatic weeds due to abundant

water availability and nutrient inflows. These conditions result in high weed biomass, as observed in the dataset collected during the December 2022 sampling campaign. Conducting the sampling in a different month, such as during the dry season, would yield data representing reduced biomass levels, as aquatic weeds experience slower growth or dieback due to lower water levels and limited nutrient availability. The temporal variability in aquatic weed biomass across seasons would therefore impact the model inputs and outputs in the following ways:

- **Model Predictions:** Seasonal variations would likely result in models trained on data from different months providing predictions reflective of the specific growth conditions for that period. For example, sampling during a dry season might yield lower biomass estimates and weaker correlations with some environmental covariates, such as precipitation indices or water availability indicators.
- **Environmental Covariates' Importance:** The relative importance of environmental covariates used in the model would change. Variables like precipitation indices, temperature, and vegetation indices (e.g., NDVI, EVI) would show season-specific variations, influencing their role as predictors. For instance, vegetation indices may have higher predictive power during active growth periods, such as December, compared to periods of senescence.
- **Yield Estimation Accuracy:** A dataset collected in a different month might reduce the accuracy of yield estimations for the months not represented in the dataset. The Gradient Boosted Regression Trees (BRT) model, which performed best in this study, might not generalize as effectively across all months without incorporating data reflecting the full range of seasonal conditions.
- **Application to Biochar Production:** Seasonal sampling differences could influence the planning and timing of biochar production. For example, a December-based dataset highlights peak harvesting opportunities, while data from other months could assist in planning for sustained harvesting schedules throughout the year.

To mitigate the effects of temporal variability, it would be advisable to conduct sampling campaigns across multiple months. This would capture a comprehensive picture of the seasonal dynamics in aquatic weed biomass, enhancing the robustness and generalizability of the machine learning models. By including data from dry and wet seasons, the models could provide more accurate and temporally flexible yield estimates for the Bangweulu Wetlands.

3.6 Implications and Applications

The findings from this study have significant practical applications and policy implications for wetland management and biochar production strategies in the Bangweulu Wetlands. The spatial yield maps generated through machine learning provide a valuable tool for decision-making in several areas:

1. Biochar Production Strategies:

- The yield maps identify high-biomass zones, allowing for targeted harvesting of aquatic weeds. This ensures the efficient allocation of resources and maximizes the biomass available for biochar production.
- Seasonal yield variation maps can guide the scheduling of harvesting activities to align with peak biomass availability, improving the sustainability of biochar supply chains.

2. Wetland Ecosystem Management:

- Controlled harvesting of aquatic weeds, informed by the yield maps, can help mitigate the negative ecological impacts of weed overgrowth, such as reduced biodiversity and water flow obstruction.
- The maps can support conservation efforts by identifying areas where weed growth poses a risk to critical habitats or wetland functions.

3. Policy and Community Applications:

- Policymakers can leverage these findings to design incentive programs that promote the use of biochar as an alternative cooking fuel, reducing dependency on wood fuel and mitigating deforestation.
- Community-based initiatives can use the yield maps to establish cooperative harvesting programs, creating employment opportunities and supporting local economies.

4. Broader Environmental Impacts:

- The integration of biochar into cooking practices reduces greenhouse gas emissions compared to traditional wood fuel, contributing to climate change mitigation efforts.
- Utilizing aquatic weeds for biochar production addresses waste management challenges while providing a renewable energy source.

These applications highlight the transformative potential of combining machine learning techniques with practical environmental management and energy production strategies. By aligning technological advancements with ecological and socioeconomic priorities, the study provides a pathway for sustainable resource use in the Bangweulu Wetlands and similar ecosystems.

4. Conclusions and Recommendations

Based on the results and analysis conducted in this study on machine learning for spatially quantifying aquatic weed yield in the Bangweulu Wetlands for biochar production as an alternative cooking fuel, the following conclusions can be drawn:

Model Performance: Among the four evaluated ML models, the Gradient-Boosted Regression Trees (BRT) model exhibited the best performance in terms of coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). It achieved an R^2 value of 0.63, indicating that it explained 63% of the variance in aquatic weed yield. The BRT model also had the lowest MAE (0.08) and RMSE (0.29), suggesting better accuracy and fit compared to the other models.

Important Predictors: The analysis of the random forest model revealed the most important predictors for estimating aquatic biomass weed yield on a dry basis in the Bangweulu Wetlands. Remote sensing indices such as the Leaf Area Index (LAI_4_DEC), Enhanced Vegetation Index (EVI_1_DEC), and Normalized Vegetation Index ($NDVI_1$_DEC) were identified as highly influential variables. Climate variables, including Mean Monthly Surface Reflectance ($M_{12}MOD_4$), Mean Temperature of Coldest Quarter (BIO_{11}), Precipitation Seasonality (BIO_{15}), and Isothermality (BIO_3), also played significant roles. Additionally, Digital Elevation Model (DEM) derivatives like Local Upslope Curvature (CRU) and Downslope Curvature (CRV) contributed to the estimation, albeit to a lesser extent.

Aquatic Biomass Weed Yield Mapping: The best-performing model, BRT, was employed to estimate the aquatic biomass weed yield on a dry basis for the entire Bangweulu Wetlands. This estimation led to the development of an aquatic biomass weed yield map (Figure 6).

Seasonal Variation: The estimated yields varied across different months, with average values ranging from 0.70 kg to 1.18 kg. These estimates were compared to the measured aquatic biomass weed yield, providing insights into the accuracy of the model predictions.

Based on the conclusions drawn from this study, the following recommendations are made:

Validation and Field Measurements: To enhance the reliability and accuracy of the ML models, it is advisable to conduct more field measurements and validation exercises. Comparing the model predictions with actual field data on aquatic biomass weed yield would provide a more robust assessment of model performance and improve the accuracy of the estimated yields.

Additional Covariates: While the selected covariates (remote sensing indices, climate variables, and DEM derivatives) proved to be important predictors, it might be beneficial to explore the inclusion of additional variables. Considering other environmental factors or incorporating socioeconomic variables could potentially improve the accuracy of the ML models for estimating aquatic weed yield.

Long-Term Monitoring: Conducting long-term monitoring of aquatic weed yield and its spatial distribution would enable the evaluation of temporal trends and changes over time. Continuous data collection would facilitate the refinement and updating of the ML models, ensuring their applicability and effectiveness in the dynamic wetland environment.

Transferability and Generalizability: As with any ML study, it is essential to assess the transferability and generalizability of the developed models. Evaluating their performance in other wetland systems or regions with similar characteristics would provide insights into the models' robustness and applicability beyond the Bangweulu Wetlands.

Integration with Decision Support Systems: Integrating the ML models with decision support systems can aid policymakers, researchers, and local communities in making informed decisions regarding biochar production and alternative cooking fuel strategies. The developed aquatic biomass weed yield map and associated models can contribute to sustainable resource management and planning in the wetland ecosystem. By implementing these recommendations, further advancements can be made in the field of machine learning for quantifying aquatic weed yield, supporting biochar production, and promoting sustainable energy solutions in wetland environments.

Acknowledgments

We extend our appreciation to the National Science and Technology Council (NSTC) of Zambia and the Technology Development & Advisory Unit (TDAU) of the University of Zambia for providing financial support for this research. Their investment in this project enabled the acquisition of necessary resources and facilitated the dissemination of findings. We are truly grateful for the collective efforts of all individuals and organizations involved in this research project. Without their support and contributions, this study would not have been possible.

Author Contributions

F.B.: conceptualization; data curation; formal analysis; investigation; methodology; software; validation; visualization; writing—original draft; writing—review and editing. L.S.: conceptualization; funding acquisition; methodology; supervision; validation; writing—review and editing. Dr. M.K.: methodology; validation; writing—review and editing. Dr. K.M.: writing—review and editing.

Competing Interests

The authors have declared that no competing interests exist.

References

1. Ekouevi K, Tuntivate V. Household energy access for cooking and heating: Lessons learned and the way forward. Washington, D.C.: World Bank Publications; 2012.
2. Kaoma M, Gheewala SH. Techno-economic assessment of bioenergy options using crop and forest residues for non-electrified rural growth centres in Zambia. Biomass Bioenergy. 2021; 145: 105944.
3. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990-2019: Update from the GBD 2019 study. J Am Coll Cardiol. 2020; 76: 2982-3021.
4. Ali J, Khan W. Factors affecting access to clean cooking fuel among rural households in India during COVID-19 pandemic. Energy Sustain Dev. 2022; 67: 102-111.
5. Gioda A. Residential fuelwood consumption in Brazil: Environmental and social implications. Biomass Bioenergy. 2019; 120: 367-375.
6. Kaoma M, Gheewala SH. Sustainability performance of lignocellulosic biomass-to-bioenergy supply chains for rural growth centres in Zambia. Sustain Prod Consum. 2021; 28: 1343-1365.
7. Li Y, Gupta R, You S. Machine learning assisted prediction of biochar yield and composition via pyrolysis of biomass. Bioresour Technol. 2022; 359: 127511.
8. IEA. Defining energy access: 2020 methodology [Internet]. Paris, France: International Energy Agency; 2020. Available from: https://policycommons.net/artifacts/1427450/defining-energy-access/2042191/.
9. Luzi L, Lin Y, Koo B, Rysankova D, Portale E. Zambia - Beyond Connections: Energy Access Diagnostic Report Based on the Multi-Tier Framework (Inglês) [Internet]. Washington, D.C.: World Bank Group; 2019. Available from: https://documents.worldbank.org/pt/publication/documents-reports/documentdetail/477041572269756712/zambia-beyond-connections-energy-access-diagnostic-report-based-on-the-multi-tier-framework.
10. Ang'u C, Muthama NJ, Mutuku MA, M'IKiugu MH. Household air pollution and its impact on human health: The case of Vihiga County, Kenya. Air Qual Atmos Health. 2022; 15: 2255-2268.
11. Lewis JJ, Hollingsworth JW, Chartier RT, Cooper EM, Foster WM, Gomes GL, et al. Biogas stoves reduce firewood use, household air pollution, and hospital visits in Odisha, India. Environ Sci Technol. 2017; 51: 560-569.
12. Pilishvili T, Loo JD, Schrag S, Stanistreet D, Christensen B, Yip F, et al. Effectiveness of six improved cookstoves in reducing household air pollution and their acceptability in rural Western Kenya. PLoS One. 2016; 11: e0165529.
13. Kaoma M, Gheewala SH. Evaluation of the enabling environment for the sustainable development of rural-based bioenergy systems in Zambia. Energy Policy. 2021; 154: 112337.
14. Shane A, Gheewala SH. Missed environmental benefits of biogas production in Zambia. J Clean Prod. 2017; 142: 1200-1209.

15. Azwar E, Mahari WAW, Liew RK, Ramlee MZ, Verma M, Chong WWF, et al. Remediation and recovery of Kariba weed as emerging contaminant in freshwater and shellfish aquaculture system via solvothermal liquefaction. Sci Total Environ. 2023; 876: 162673.
16. Carregosa ISC, de Carvalho Carregosa J, Silva WR, Santos TM, Wisniewski Jr A. Thermochemical conversion of aquatic weed biomass in a rotary kiln reactor for production of bio-based derivatives. J Anal Appl Pyrolysis. 2023; 173: 106048.
17. Kaur M, Kumar M, Sachdeva S, Puri S. Aquatic weeds as the next generation feedstock for sustainable bioenergy production. Bioresour Technol. 2018; 251: 390-402.
18. Kumari K, Swain AA, Kumar M, Bauddh K. Utilization of Eichhornia crassipes biomass for production of biochar and its feasibility in agroecosystems: A review. Environ Sustain. 2021; 4: 285-297.
19. Song H, Wang J, Garg A, Lin X, Zheng Q, Sharma S. Potential of novel biochars produced from invasive aquatic species outside food chain in removing ammonium nitrogen: Comparison with conventional biochars and clinoptilolite. Sustainability. 2019; 11: 7136.
20. Kalumba M, Nyirenda E, Nyambe I, Dondeyne S, Van Orshoven J. Machine learning techniques for estimating hydraulic properties of the topsoil across the Zambezi River Basin. Land. 2022; 11: 591.
21. Meshram V, Patil K, Meshram V, Hanchate D, Ramkteke S. Machine learning in agriculture domain: A state-of-art survey. Artif Intell Life Sci. 2021; 1: 100010.
22. Dube T, Mutanga O, Adam E, Ismail R. Intra-and-inter species biomass prediction in a plantation forest: Testing the utility of high spatial resolution spaceborne multispectral rapideye sensor and advanced machine learning algorithms. Sensors. 2014; 14: 15348-15370.
23. Greenhall J, Pantea C, Vakhlamov P, Davis E, Semelsberger T. Data-driven acoustic measurement of moisture content in flowing biomass. Mach Learn Appl. 2023; 13: 100476.
24. Tamiminia H, Salehi B, Mahdianpari M, Beier CM, Johnson L, Phoenix DB. A comparison of decision tree-based models for forest above-ground biomass estimation using a combination of airborne lidar and landsat data. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci. 2021; 3: 235-241.
25. Wang X, Xu G, Feng Y, Peng J, Gao Y, Li J, et al. Estimation model of rice aboveground dry biomass based on the machine learning and hyperspectral characteristic parameters of the canopy. Agronomy. 2023; 13: 1940.
26. Li H, Ai X, Wang L, Zhang R. Substitution strategies for cooking energy: To use gas or electricity? J Environ Manage. 2022; 303: 114135.
27. Zhu X, Li Y, Wang X. Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Bioresour Technol. 2019; 288: 121527.
28. Karimi Y, Prasher S, Patel R, Kim S. Application of support vector machine technology for weed and nitrogen stress detection in corn. Comput Electron Agric. 2006; 51: 99-109.
29. Liu B, Li R, Li H, You G, Yan S, Tong Q. Crop/weed discrimination using a field imaging spectrometer system. Sensors. 2019; 19: 5154.
30. Schmitter P, Steinrücken J, Römer C, Ballvora A, Léon J, Rascher U, et al. Unsupervised domain adaptation for early detection of drought stress in hyperspectral images. ISPRS J Photogramm Remote Sens. 2017; 131: 65-76.

31. Sujud L, Jaafar H, Hassan MAH, Zurayk R. Cannabis detection from optical and RADAR data fusion: A comparative analysis of the SMILE machine learning algorithms in Google Earth Engine. Remote Sens Appl. 2021; 24: 100639.
32. Wang P, Tan S, Zhang G, Wang S, Wu X. Remote sensing estimation of forest aboveground biomass based on Lasso-SVR. Forests. 2022; 13: 1597.
33. Bakhshipour A, Jafari A. Evaluation of support vector machine and artificial neural networks in weed detection using shape features. Comput Electron Agric. 2018; 145: 153-160.
34. Pereira LA, Nakamura RY, De Souza GF, Martins D, Papa JP. Aquatic weed automatic classification using machine learning techniques. Comput Electron Agric. 2012; 87: 56-63.
35. Tao T, Wei X. A hybrid CNN-SVM classifier for weed recognition in winter rape field. Plant Methods. 2022; 18: 29.
36. Ayyash F, Hayslep M, Ko T, Kalumba M, Simukonda K, Farmani R. Application of a neural network model to short-term water demand forecasting. Eng Proc. 2024; 69: 123.
37. Muloiwa M, Dinka M, Nyende-Byakika S. Application of artificial neural network for predicting biomass growth during domestic wastewater treatment through a biological process. Eng Life Sci. 2023; 23: e2200058.
38. Peng W, Karimi Sadaghiani O. Machine learning for sustainable reutilization of waste materials as energy sources-a comprehensive review. Int J Green Energy. 2024; 21: 1641-1666.
39. Khan M, Ullah Z, Mašek O, Naqvi SR, Khan MNA. Artificial neural networks for the prediction of biochar yield: A comparative study of metaheuristic algorithms. Bioresour Technol. 2022; 355: 127215.
40. Ali ZA, Abduljabbar ZH, Tahir HA, Sallow AB, Almufti SM. eXtreme gradient boosting algorithm with machine learning: A review. Acad J Nawroz Univ. 2023; 12: 320-334.
41. Herrero-Huerta M, Rodriguez-Gonzalvez P, Rainey KM. Yield prediction by machine learning from UAS-based multi-sensor data fusion in soybean. Plant Methods. 2020; 16: 1-16.
42. Montomoli J, Romeo L, Moccia S, Bernardini M, Migliorelli L, Berardini D, et al. Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. J Intensive Med. 2021; 1: 110-116.
43. Nininahazwe F, Théau J, Marc Antoine G, Varin M. Mapping invasive alien plant species with very high spatial resolution and multi-date satellite imagery using object-based and machine learning techniques: A comparative study. GIsci Remote Sens. 2023; 60: 2190203.
44. Pathy A, Meher S. Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. Algal Res. 2020; 50: 102006.
45. Ben Ayed R, Hanana M. Artificial intelligence to improve the food and agriculture sector. J Food Qual. 2021; 2021: 5584754.
46. Myneni R, Knyazikhin Y, Park T. MOD15A2H MODIS/Terra leaf area Index/FPAR 8-Day L4 global 500 m SIN grid V006 [Internet]. Sioux Falls, SD: LP DAAC; 2015. Available from: https://lpdaac.usgs.gov/products/mod15a2hv006/.
47. Kalumba M, Dondeyne S, Vanuytrecht E, Nyirenda E, Van Orshoven J. Functional evaluation of digital soil hydraulic property maps through comparison of simulated and remotely sensed maize canopy cover. Land. 2022; 11: 618.
48. Fick SE, Hijmans RJ. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. Int J Climatol. 2017; 37: 4302-4315.

49. Darwall W, Smith K, Allen D, Holland R, Harrison I, Brooks E. The diversity of life in African freshwaters: Underwater, under threat: An analysis of the status and distribution of freshwater species throughout mainland Africa. Gland, Switzerland: International Union for Conservation of Nature; 2011.
50. Nussbaum M, Spiess K, Baltensweiler A, Grob U, Keller A, Greiner L, et al. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil. 2018; 4: 1-22.
51. Samuel-Rosa A, Heuvelink G, Vasques G, Anjos L. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma. 2015; 243-244: 214-227.
52. Baboo SS, Devi MR. An analysis of different resampling methods in Coimbatore, District. Global J Comput Sci Technol. 2010; 10: 61-66.
53. Zomer RJ, Bossio DA, Trabucco A, Yuanjie L, Gupta DC, Singh VP. Trees and water: Smallholder agroforestry on irrigated lands in Northern India. Colombo, Sri Lanka: IWMI; 2007.
54. Zomer RJ, Trabucco A, Bossio DA, Verchot LV. Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agric Ecosyst Environ. 2008; 126: 67-80.
55. Didan K. MODIS/Terra vegetation indices 16-day L3 global 250 m SIN grid V061 [Internet]. Sioux Falls, SD: LP DAAC; 2021. Available from: https://lpdaac.usgs.gov/products/mod13q1v061/.
56. Liu J, Yue C, Pei C, Li X, Zhang Q. Prediction of regional forest biomass using machine learning: A case study of Beijing, China. Forests. 2023; 14: 1008.
57. Michelini D, Dalponte M, Carriero A, Kutchartt E, Pappalardo SE, De Marchi M, et al. Hyperspectral and LiDAR data for the prediction via machine learning of tree species, volume and biomass: A contribution for updating forest management plans. In: Geomatics for Green and Digital Transition. Cham: Springer; 2022. pp. 235-250.
58. Huy B, Truong NQ, Khiem NQ, Poudel KP, Temesgen H. Deep learning models for improved reliability of tree aboveground biomass prediction in the tropical evergreen broadleaf forests. For Ecol Manage. 2022; 508: 120031.
59. Ferdous SN, Li X, Sahoo K, Bergman R. Analysis of biomass sustainability indicators from a machine learning perspective. arXiv. 2023. doi: 10.48550/arXiv.2302.00828.
60. Chan T, Gomez CA, Kothikar A, Baiz P. Joint study of above ground biomass and soil organic carbon for total carbon estimation using satellite imagery in scotland. arXiv. 2022. doi: 10.48550/arXiv.2205.04870.
61. Earth Resources Observation and Science (EROS) Center. USGS EROS Archive - Digital Elevation - Shuttle Radar Topography Mission (SRTM) Void Filled [Internet]. Sioux Falls, SD: Earth Resources Observation and Science (EROS) Center; 2018. Available from: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-void?qt-science_center_objects=0#qt-science_center_objects.
62. Kuhn M, Johnson K. Applied predictive modeling. New York, NY: Springer; 2013.
63. Hlatshwayo ST, Mutanga O, Lottering RT, Kiala Z, Ismail R. Mapping forest aboveground biomass in the reforested Buffelsdraai landfill site using texture combinations computed from SPOT-6 pan-sharpened imagery. Int J Appl Earth Obs Geoinf. 2019; 74: 65-77.
64. Kumar Y, Babu S, Singh S. The potential of radar remote sensing in estimation and mapping of forest above ground biomass in sub-tropical forest (Kempti Forest Range) in the Himalayas. Environ Ecol. 2020; 38: 715-724.

65. Vorster AG, Evangelista PH, Stovall AE, Ex S. Variability and uncertainty in forest biomass estimates from the tree to landscape scale: The role of allometric equations. Carbon Balance Manage. 2020; 15: 8.
66. Galidaki G, Zianis D, Gitas I, Radoglou K, Karathanassi V, Tsakiri-Strati M, et al. Vegetation biomass estimation with remote sensing: Focus on forest and other wooded land over the Mediterranean ecosystem. Int J Remote Sens. 2017; 38: 1940-1966.
67. Reddersen B, Fricke T, Wachendorf M. A multi-sensor approach for predicting biomass of extensively managed grassland. Comput Electron Agric. 2014; 109: 247-260.
68. Xue J, Su B. Significant remote sensing vegetation indices: A review of developments and applications. J Sens. 2017; 2017: 1353691.
69. Mahgoub AM. The impact of five environmental factors on species distribution and weed community structure in the coastal farmland and adjacent territories in the northwest delta region, Egypt. Heliyon. 2019; 5: e01441.
70. Peters K, Breitsameter L, Gerowitt B. Impact of climate change on weeds in agriculture: A review. Agron Sustain Dev. 2014; 34: 707-721.
71. Ramesh K, Matloob A, Aslam F, Florentine SK, Chauhan BS. Weeds in a changing climate: Vulnerabilities, consequences, and implications for future weed management. Front Plant Sci. 2017; 8: 95.
72. Dubeau P, King DJ, Unbushe DG, Rebelo LM. Mapping the Dabus wetlands, ethiopia, using random forest classification of Landsat, PALSAR and topographic data. Remote Sens. 2017; 9: 1056.
73. Gxokwe S, Dube T, Mazvimavi D. An assessment of long-term and large-scale wetlands change dynamics in the Limpopo transboundary river basin using cloud-based Earth observation data. Wetlands Ecol Manage. 2024; 32: 89-108.
74. O'Neil GL, Goodall JL, Watson LT. Evaluating the potential for site-specific modification of LiDAR DEM derivatives to improve environmental planning-scale wetland identification using Random Forest classification. J Hydrol. 2018; 559: 192-208.
75. Yu H, Cooper AR, Infante DM. Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. Ecol Modell. 2020; 432: 109202.
76. Qiu X, Zhang L, Suganthan PN, Amaratunga GA. Oblique random forest ensemble via least square estimation for time series forecasting. Inf Sci. 2017; 420: 249-262.
77. Vergni L, Todisco F. A random forest machine learning approach for the identification and quantification of erosive events. Water. 2023; 15: 2225.