Identification and Removal of Negative Biomass Samples via Scatter Plot Analysis to Improve GWP Predictive Modeling

Prakash Gyawali[1], *Bijendra* Shrestha[2], *Jetsada* Posom[3,*], *Pimpen* Pornchaloempong[4], *Panmanas* Sirisomboon[1], *Bim Prasad* Shrestha[2,5,*], and *Axel* Funke[6]

[1]Department of Agricultural Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand, 66016181@kmitl.ac.th.
[2]Department of Mechanical Engineering, School of Engineering, Kathmandu University, Dhulikhel, PO Box 6250, Nepal, 63601254@kmitl.ac.th
[3]Department of Agricultural Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand, jetspo@kku.ac.th
[4]Department of Food Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand, pimpen.po@kmitl.ac.th
[5]Department of BioEngineering, University of Washington, Seattle, William H. Foege Building 3720, 15th Ave NE, Seattle, WA 98195-5061, USA, shrestha@ku.edu.np
[6]Institute of Catalysis Research and Technology (IKFT), Karlsruhe Institute of Technology (KIT), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany, axel.funke@kit.edu

Abstract. Accurate prediction of Global Warming Potential (GWP) from biomass constituents is essential for evaluating the sustainability of bioenergy sources. However, the inclusion of biomass samples with weak or negative correlation to key elemental components such as Carbon (C), Hydrogen (H), Nitrogen (N), and Oxygen (O)—can reduce model accuracy and lead to misleading conclusions. This study utilizes scatter plot regression analysis to evaluate and remove "negative biomass samples," defined as those with consistently low R^2 values across constituent-GWP relationships using HHV = 0.2949C + 0.8250H developed for wood biomass by Yin. Regression models were generated for each biomass species using elemental concentrations as predictors of GWP. Notably, several non-wood species (e.g., Zea Mays-Shell, Bagasse, Bamboo) exhibited very low R^2 values (often <0.05) for model between elemental composition and GWP, where all elemental correlations indicated weak predictive relationships. In contrast, wood-based species such as Alnus demonstrated significantly higher R^2 values, especially with Carbon (R^2 = 0.69), Hydrogen (R^2 = 0.57), and Oxygen (R^2 = 0.68), suggesting a stronger linear influence on GWP. Removing these low-contributing samples improved the clarity and reliability of the predictive model related to HHV and each type of element (C,H,N and O) as evidenced by a sharper regression slope of a graph plotted between predicted GWP and measured GWP of positive species and better fit (increased R^2) for the remaining samples. These results highlight the value of preliminary scatter plot analysis in identifying biomass species that obscure rather than support predictive modeling. This filtering step ultimately enhances the robustness and interpretability of constituent-based GWP prediction frameworks, particularly when applying FT-NIR spectroscopy and chemometric modelling.

Keywords: Biomass, GWP, Scatter Plot, Regression, Ultimate Analysis, Model Optimization

* Corresponding author: jetspo@kku.ac.th and shrestha@ku.edu.np

1 Introduction

Global Warming Potential (GWP) is a standardized metric introduced by the Intergovernmental Panel on Climate Change (IPCC) to compare the climate impact of greenhouse gases based on their radiative forcing over a specific time horizon[1]. Biomass combustion is widely promoted as a renewable alternative to fossil fuels, but its environmental footprint varies considerably depending on the feedstock composition and combustion efficiency[2]. Therefore, accurate prediction of GWP for different biomass species is crucial for informed climate policy, sustainable energy planning, and lifecycle assessment. Two primary forms of biomass characterization are commonly employed in combustion modeling : ultimate analysis and proximate analysis. Ultimate analysis quantifies elemental components such as carbon (C), hydrogen (H), nitrogen (N), sulfur (S), and oxygen (O), while proximate analysis provides values for volatile matter (VM), fixed carbon (FC), moisture content (MC), and ash [3]. These parameters directly influence combustion behavior and emissions, thereby affecting the overall GWP. Recent studies have demonstrated the use of empirical and statistical models to estimate GWP from such data[4]. However, the reliability of these models can be hindered by the presence of anomalous data or species with atypical combustion properties. In this context, scatter plot analysis emerges as a powerful visual tool for identifying inconsistencies and guiding the refinement of predictive models [5]. This study proposes a scatter plot-based methodology for improving GWP modeling accuracy by detecting and removing negative or outlier biomass samples. By analyzing the linear relationship between individual constituents (e.g., %C, %H, %N, %O) and GWP values, we demonstrate that model calibration metrics such as R^2, slope, and intercept are significantly improved following the exclusion of problematic samples. The proposed approach enables better species-specific calibration and highlights the importance of exploratory data visualization in environmental modeling.

1.1 The Role of Data Quality in Regression Modeling

Data quality plays a foundational role in the effectiveness and interpretability of regression modeling. High-quality data ensures that the relationships identified through modeling are both statistically valid and practically meaningful. Poor data quality—manifested through missing values, outliers, measurement errors, or incorrect variable types—can distort regression coefficients, inflate error terms, and reduce model generalizability [6]. Regression models are particularly sensitive to anomalies such as outliers and incorrect data entries. These issues can disproportionately influence parameter estimates, especially in small sample sizes or in models with multicollinearity [7]. For example, negative or biologically implausible values in ecological data, such as negative biomass, can severely bias predictions of related variables such as Global Warming Potential (GWP), unless appropriately identified and removed through pre-modeling diagnostics. Moreover, ensuring data consistency and accuracy facilitates better feature selection, improves the performance of regularization techniques, and supports the validity of inferential statistics [8]. This is especially important in environmental modeling, where data may come from diverse sources with varying levels of precision and calibration. Therefore, a rigorous data cleaning process—including scatter plot diagnostics, anomaly detection, and metadata validation—is essential before applying any regression technique. As modern machine learning and statistical modeling techniques grow more sophisticated, the importance of high-quality input data remains constant, serving as the bedrock for all predictive accuracy and interpretive trustworthiness [9]

2 Methodology

Biomass samples which were report Shrestha et al[10], were collected from Nepal's Terai low flatland and mid-hill regions, spanning altitudes from 86 to 1940 meters above sea level. The study encompassed five fast-growing species: (1) *Alnus nepalensis*, (2) *Pinux roxiburghii*, (3) *Bombusa vulagris*, (4) *Bombax ceiba*, and (5) *Eucalyptus camaldulensis*, along with five agricultural residues: (1*) Zea mays* (cob), (2) *Zea mays (shell*), (3) *Zea mays* (stover), (4) *Oryza sativa* (husk), and (5) *Saccharum officinarum* (bagass). *Alnus nepalensis* and *Pinux roxiburghii* were sourced from the mid-hill region, while *Bombax ceiba, Eucalyptus camaldulensis*, and *Saccharum officinarum* (bagass) were from the Terai region. *Zea mays* (cob, shell, stover), *Bombusa vulagris,* and *Oryza sativa* were collected from both Terai and mid-hill regions of Nepal. All samples, except *Oryza sativa*, were manually chopped into pieces smaller than 30 mm x 15 mm, sun-dried, and stored in airtight aluminum bags to preserve their biomass properties. These samples were transported to the Near-Infrared Spectroscopy Research Center for Agricultural Product and Food at the School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Thailand.

In this study, MATLAB-R2020b (MathWorks, NC, USA) was utilized to create PLSR calibration models using spectral data from ten distinct biomass varieties (including five fast-growing tree varieties and five agricultural residue varieties). All chip biomass samples were scanned using an FT-NIR spectrometer (MPA, Bruker, Ettlingen, Germany) in diffuse reflectance with sphere macro sample rotating mode, covering the wavenumber range from 3594.87 to 12,489 cm^{-1}, with a resolution of 16 cm^{-1}[10]. The scanning process consisted of 32 scans (on average) for both sample and background scans to collect the raw spectra in absorbance mode of log (1/R). R is the diffuse reflectance detected from the chipped biomass sample. Prior to scanning, the FT-NIR spectrometer was performed for background compensation by a gold plate internal scan. The primary purpose of performing a background scan on every new sample was to

compensate for instrument signal drift from ambient environmental influences, such as temperature, relative humidity, etc. on the measurement setup

Reference data were obtained from [10] a bomb calorimeter for HHV (TJ/kg); a CHNS elemental analyzer for carbon (C), nitrogen (N), hydrogen (H), sulfur (S) and Oxygen (O) content and thermogravimetric analyzer for ash content in weight percent and the GWP determined by using IPCC guidelines for stationary biomass

The finding of GWP (Dependent variable) using HHV = 0.2949C + 0.8250H where are Independent variable is CHNS elemenment analyzer data. The total GWP calculates the combined impact of different greenhouse gases (GHGs) on global warming. Each GHG has a specific GWP value, which represents its warming effect relative to carbon dioxide (CO_2) over a specific time horizon, typically 100 years. For calculation, the total GWP is determined by summing the products of the GWP values and the emissions quantities for each gas, in our case; carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) Specifically,

GWP total = (1×CO_2 Emission) + (29.8×CH_4 Emission) + (273×N_2O Emission) for 100 year (1)

GHG Emissions (kg) = Mass of biomass sample (kg) × HHV of biomass sample (TJ/kg) × Emission Factor (EF) of corresponding GHG (kg/TJ)

By accounting for the different contributions of these gases, the formula provides a comprehensive measure of the overall impact of multiple GHGs on climate change, allowing for a more accurate assessment of their collective influence on global warming.

2.1 Linear Regression Analysis

$$y = mx + b \quad (2)$$

Where: y is the predicted Global Warming Potential (GWP), x is the elemental variable (such as %Carbon), m shows how much GWP increases or decreases with each unit change in x and b gives the baseline GWP when x = 0

2.2 Slope (m) and Intercept (b) calculation

Given n samples (xi, yi), the slope and intercept are calculated as :

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$b = \bar{y} - m\bar{x} \quad (3)$$

Where $\bar{x}$ and $\bar{y}$ are the means of xi and yi respectively

m = Slope, representing the rate of change in GWP with respect to the constituent

b = Intercept representing the predicted GWP when x=0

Coefficient of Determination (R^2)

The coefficient of determination indicates how well the regression model explains the variability of GWP. It is computed as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

y_i = Actual GWP value

$\hat{y}_i$ = Predicted GWP from Regression

$\bar{y}$ = GWP mean

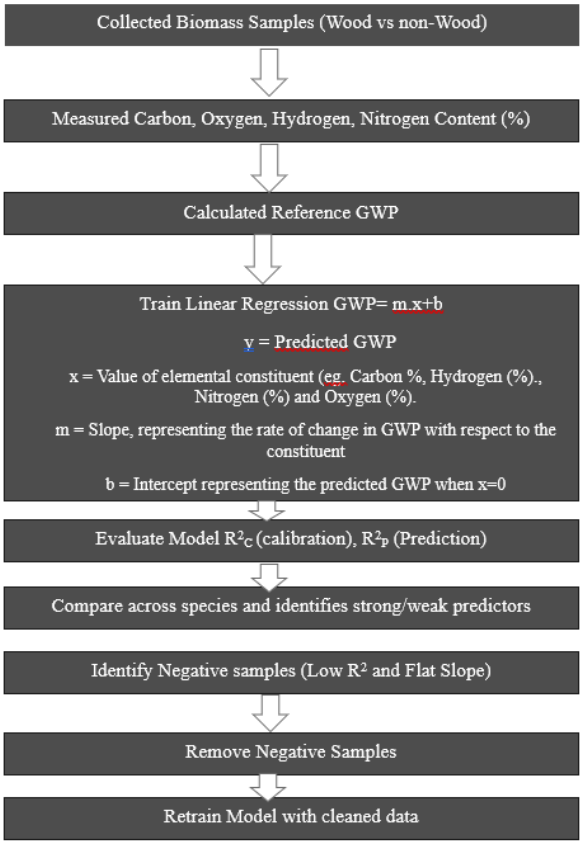

Fig. 1. The model of the Scatter Plot optimized model for a) C, b) H, c) O and d) N where the simple regression lines of measured and predicted to improve GWP Predictive Modeling Result.

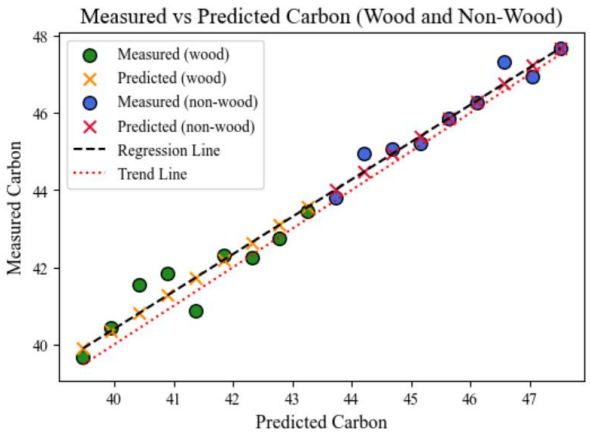

Fig. 2. The scatter plots depict the relationship between the measured carbon content and the values predicted by the NIR model.

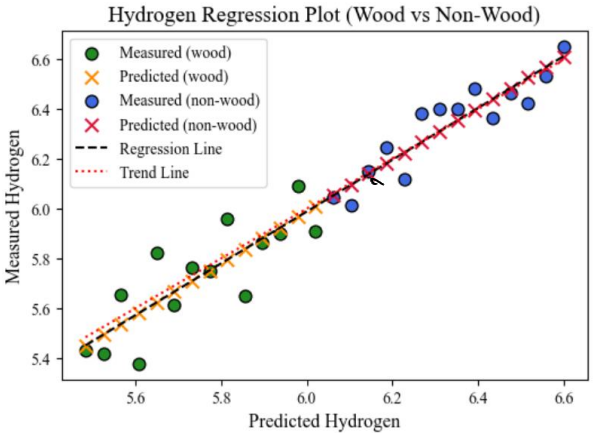

Fig. 3. The scatter plots depict the relationship between the measured hydrogen content and the values predicted by the NIR model.

Fig 3 graph presents a linear regression analysis demonstrating a strong correlation between predicted and measured hydrogen content for both wood and non-wood samples. The data points from both material types are tightly aligned with the regression and trendlines, which run nearly identically through the center of the data. This high degree of linearity and the consistent positive slope of the trendline indicate that the predictive model is highly effective and accurate. The fact that a single regression line fits both the wood (green circles) and non-wood (blue circles) data equally well suggests that the predictive model is robust and applicable across both material types without significant performance deviation.

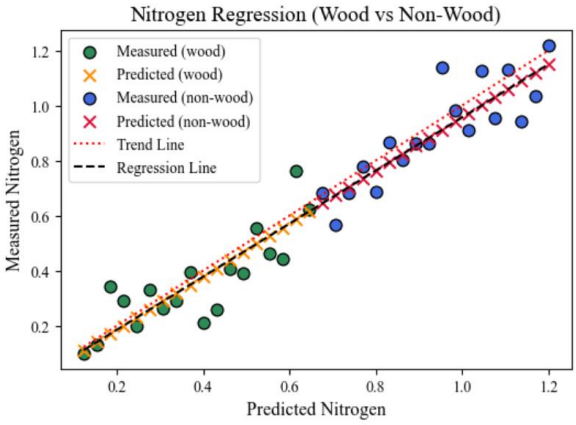

Fig. 4. The scatter plots depict the relationship between the measured Nitrogen content and the values predicted by the NIR model.

Fig 4 illustrates a strong, positive linear correlation between the predicted and measured nitrogen content for both wood and non-wood samples. The regression and trendlines, which are nearly identical, effectively model the relationship, indicating that as the predicted nitrogen content increases, the measured content also increases proportionally. The fact that the data for both material types (green and blue circles) aligns closely with a single trendline suggests that the predictive model is robust and can be applied consistently to both wood and non-wood materials without the need for separate adjustments.

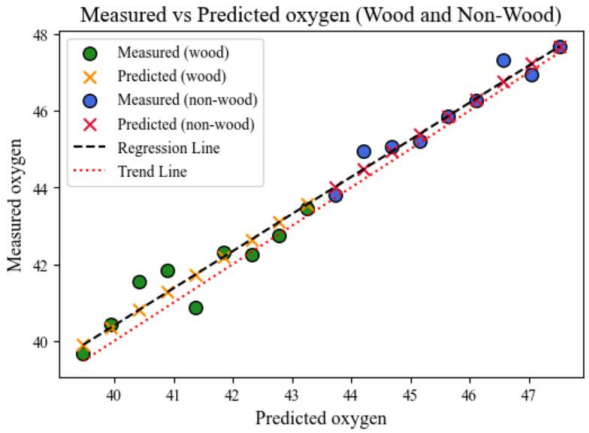

Fig. 5. The scatter plots depict the relationship between the measured oxygen content and the values predicted by the NIR model.

This regression plot shows a strong linear correlation between the predicted and measured oxygen content for both wood and non-wood samples. The data points from both materials are closely clustered around a single regression line, which has a positive slope. This indicates that as the predicted oxygen value increases, the measured oxygen value also increases predictably. The high degree of alignment suggests that the predictive model is accurate and works well for both types of materials.

Table 1 The trend line charateristic of the wood and non wood species in scatter plots of the best models for C, H, N and O

Feature	R^2_C	R^2_P	$Slope_{Cn}$	$Slope_P$	$Intercept_{Cn}$	$Intercept_P$
C	0.1110	0.9498	-0.000185	0.3864	0.043941	0.6671
H	0.1879	0.7314	-0.001554	2.1936	0.044648	5.2726
N	0.0077	0.0340	-0.000383	-0.5761	0.035843	18.0739
O	0.1414	0.9808	0.000190	-0.3589	0.026318	35.6006

R2c : coefficient of determination (R^2) in the calibration set, R2p :Coefficient of determination in the prediction set, SlopeC : Slope of trendline in the calibration set, Slopep : Slope of trendline in the prediction set, Interceptc : Intercept in the calibration set, Interceptp : Intercept in the prediction set

The table 1 illustrate The element of C, H, N, and S on a dry basis in the chip biomass data were determined at the Scientific and Technological Research Equipment Center (STREC) at Chulalongkorn University, Bangkok, Thailand, using CHNS/O analyzer (Thermo Scientific™ FLASH 2000, Waltham, MA, USA)[10]

Based on the trend line characteristics from the calibration and prediction sets for C, H, N, and O, the model performances varied notably between wood and non-wood biomass species. For carbon, the prediction performance was strong (R^2p= 0.9498), mainly due to the non-wood samples, but the $Slope_C$ (0.3864) and intercept (0.6671) deviated from ideal values, while the calibration performance was poor (R^2_C = 0.1110). The hydrogen model showed moderate prediction strength (R^2_P = 0.7314), but an oversteep slope (2.1936) and high intercept (5.2726) suggested mismatched trends between sample types. Nitrogen showed the weakest model performance, with very low R^2 values in both calibration (0.0077) and prediction (0.0340), a negative slope $_C$ (-0.5761), and a large intercept (18.0739), indicating severe overfitting and unreliable trends. For oxygen, prediction performance was excellent (R^2 = 0.9808), again driven by non-wood samples, but the slope (-0.3589) and intercept (35.6006) revealed prediction bias, while calibration remained poor (R^2 = 0.1414). Overall, non-wood samples contributed more positively to prediction accuracy, whereas wood samples showed inconsistent or poor trend characteristics across all models.

Table 2 The trend line charateristic of specific biomass species for carbone valuation optimized model

Type	Biomass Species	R^2_C	R^2_P	$Slope_{Cn}$	$Slope_P$	$Intercept_{Cn}$	$Intercept_P$
Non-Wood	Bagasse	0.0123	1	0.0123	0.0123	0.0347	0.0347
Non-Wood	Bamboo	0.1259	1	0.1259	0.1259	0.0309	0.0309
Non-Wood	Rice husk	0.1576	1	0.1576	0.1576	0.0307	0.0307
Non-Wood	*Zea Mays*-Cob	0.042	1	0.042	0.042	0.0344	0.0344
Non-Wood	*Zea Mays*-Shell	0.0287	1	0.0287	0.0287	0.0356	0.0356
Non-Wood	*Zea Mays*-Stover	0.0408	1	0.0408	0.0408	0.0338	0.0338
Wood	Alnus	0.6923	1	0.6923	0.6923	0.0112	0.0112
Wood	Bombax	0.0107	1	0.0107	0.0107	0.037	0.037
Wood	Euca	0.0352	1	0.0352	0.0352	0.0332	0.0332
Wood	Pine	0.1221	1	0.1221	0.1221	0.0302	0.0302

The table 2 indicate the trend line characteristics of specific biomass species in the optimized carbon prediction model reveal notable differences between wood and non-wood groups. Among the non-wood species, all samples achieved perfect prediction fit (R^2_P = 1), but with low slope values ranging from 0.0123 (Bagasse) to 0.1576 (Rice husk), indicating that although the model captured variance, it did so with limited proportional accuracy. Intercept values were consistently higher in non-wood species (around 0.0307–0.0356), reflecting a slight positive bias. Within the wood group, Alnus exhibited the strongest calibration performance (R^2_C = 0.6923, slope = 0.6923), indicating relatively better model fitting and trend representation, though its intercept was comparatively lower (0.0112), suggesting a closer alignment with the ideal prediction line. In contrast, other wood species such as Bombax (R^2_C = 0.0107), Eucalyptus (0.0352), and Pine (0.1221) showed weak calibration trend fitting, similar to most non-wood species. Despite the perfect R^2 in all prediction sets, the generally low slope values in both groups highlight that while the model accurately predicts carbon concentration rankings, it does not closely follow the actual reference values, especially for species with minimal trend contribution. These findings suggest that although non-wood species contribute to high apparent prediction accuracy, only select species like Alnus demonstrate meaningful calibration trend alignment, indicating its potential as a representative biomass for model training. R^2p is high here because the prediction samples match the calibration patterns, but R^2c is low because mixed-species data weakens the fundamental calibration relationship—only species with consistent chemical profiles (like *Alnus*) give strong calibration performance.

The uniformly high R^2p values (all = 1) but generally low R^2c values indicate that the prediction samples matched the model trends almost perfectly, while the calibration itself was weak. This mismatch is largely due to the mixed biomass species used for model development. NIR spectroscopy relies on detecting subtle and consistent spectral–chemical relationships. When multiple, chemically and structurally different species are combined—such as bagasse, bamboo, rice husk, and pine—the spectral variability becomes too large, and the correlation with the target variable weakens, leading to low R^2c. In contrast, when the dataset is restricted to a single species or closely related species, the chemical composition and structural properties are more uniform, allowing NIR to capture stable patterns and achieve higher calibration accuracy. For example, Alnus shows a relatively strong calibration performance (R^2c = 0.5719) because its intra-species variation is low, while pine (R^2c = 0.0056) performs poorly due to its distinct spectral profile that does not align with other species in the model

Table 3 The trend line charateristic of specific biomass species for nitrogen valuation optimized model

Type	Biomass Species	R^2_C	R^2_P	Slope C	Slope P	Intercept C	Intercept P
Non-Wood	Bagasse	0.0256	1	0.0256	0.0256	0.0342	0.0342
Non-Wood	Bamboo	0.0011	1	0.0011	0.0011	0.0353	0.0353
Non-Wood	Rice husk	0	1	0	0	0.0364	0.0364
Non-Wood	*Zea Mays*-Cob	0.0041	1	0.0041	0.0041	0.0357	0.0357
Non-Wood	*Zea Mays*-Shell	0.0021	1	0.0021	0.0021	0.0365	0.0365
Non-Wood	*Zea Mays*-Stover	0.0339	1	0.0339	0.0339	0.0341	0.0341
Wood	Alnus	0.0955	1	0.0955	0.0955	0.0328	0.0328
Wood	Bombax	0.0002	1	0.0002	0.0002	0.0374	0.0374
Wood	Euca	0.0735	1	0.0735	0.0735	0.0318	0.0318
Wood	Pine	0.0197	1	0.0197	0.0197	0.0337	0.0337

The table 3 trend line characteristics of specific biomass species in the nitrogen valuation optimized model indicate consistently poor calibration performance across both non-wood and wood types, despite a perfect prediction ($R^2_P = 1$) for all species. Among non-wood species, calibration R^2 values were extremely low, ranging from 0 (rice husk) to 0.0339 (zea mays-stover), with correspondingly minimal slopes (all < 0.0355), indicating weak correlation between measured and predicted nitrogen values. Intercepts for these species were relatively uniform (0.0341–0.0365), suggesting limited spread but consistent prediction offset. Similarly, wood species showed slightly better but still low calibration performance, with Alnus (R^2_C = 0.0955) and Eucalyptus (0.0735) exhibiting the highest R^2 values within this group. Their slope values mirrored the R^2 results, suggesting modest alignment in trend, while intercepts remained low (0.0318–0.0374). The uniformly perfect R^2_P values for all species imply that the model predicted nitrogen concentrations with high consistency across samples; however, the very low calibration slopes and R^2_C values reflect a lack of true model learning or generalization. These results highlight that, although the nitrogen model appears accurate in prediction plots, it fails to capture meaningful variation across species, particularly in non-wood biomass, thus limiting its practical utility for quantitative nitrogen estimation. The mixed-species modelling with NIR is not ideal because species-specific chemical differences disrupt calibration, whereas species-specific models—such as for *Alnus*—maintain consistent spectral–chemical relationships and yield more reliable predictions.

Table 4 The trend line charateristic of specific biomass species for hydrogen valuation optimized model

Type	Biomass Species	R^2_C	R^2_P	$Slope_C$	$Slope_P$	$Intercept_C$	$Intercept_P$
Non-Wood	Bagass	0.2675	1	0.2675	0.2675	0.0257	0.0257
Non-Wood	Bamboo	0.0865	1	0.0865	0.0865	0.0323	0.0323
Non-Wood	Ricehusk	0.1677	1	0.1677	0.1677	0.0303	0.0303
Non-Wood	*Zea Mays*-Cob	0.0077	1	0.0077	0.0077	0.0356	0.0356
Non-Wood	*Zea Mays*-Shell	0.0001	1	0.0001	0.0001	0.0366	0.0366
Non-Wood	*Zea Mays*-Stover	0.0106	1	0.0106	0.0106	0.0349	0.0349
Wood	Alnus	0.5719	1	0.5719	0.5719	0.0155	0.0155
Wood	Bombax	0.0234	1	0.0234	0.0234	0.0366	0.0366
Wood	Euca	0.0398	1	0.0398	0.0398	0.033	0.033
Wood	Pine	0.0056	1	0.0056	0.0056	0.0342	0.0342

The table 4 the trend line characteristics of specific biomass species in the hydrogen valuation optimized model show wide variability in calibration performance across both non-wood and wood types, despite a perfect prediction R^2 ($R^2_P = 1$) for all species. Among non-wood species, bagasse (R^2_C = 0.2675) and rice husk (0.1677) exhibited relatively stronger calibration fits, with moderate slope values (0.2675 and 0.1677, respectively), suggesting partial alignment between measured and predicted hydrogen content. In contrast, bamboo and all Zea mays components (cob, shell, stover) showed poor calibration ($R^2C < 0.0875$), with very low slope values (0.0001–0.0865), indicating minimal predictive reliability within the calibration set. Intercepts in the non-wood group were higher (ranging from 0.0257 to 0.0366), reflecting a consistent prediction offset. Among wood species, Alnus stood out with the best calibration performance (R^2C = 0.5719; slope = 0.5719), indicating strong correlation and trend accuracy for hydrogen prediction, while the remaining wood samples—Bombax, Eucalyptus, and Pine—had low R^2C values (≤ 0.0398) and shallow slopes, suggesting poor model learning. The relatively lower intercepts observed in wood species, especially for Alnus (0.0955), indicate less bias in predicted values. Despite the perfect R^2 in all prediction plots, the consistently low calibration metrics for most species—particularly non-wood—suggest that the model may not generalize well and might reflect overfitting. Alnus and bagasse appear to be the most reliable biomass types for hydrogen model development due to their comparatively stronger calibration performance.

Table 5 The trend line charateristic of specific biomass species for oxygen valuation optimized model

Type	Biomass Species	R^2_C	R^2_P	$Slope_C$	$Slope_P$	$Intercept_C$	$Intercept_P$
Non-Wood	Bagass	0.0472	1	0.0472	0.0472	0.0334	0.0334
Non-Wood	Bamboo	0.1539	1	0.1539	0.1539	0.0299	0.0299
Non-Wood	Ricehusk	0.1724	1	0.1724	0.1724	0.0301	0.0301
Non-Wood	*Zea Mays*-Cob	0.0362	1	0.0362	0.0362	0.0346	0.0346
Non-Wood	*Zea Mays*-Shell	0.0259	1	0.0259	0.0259	0.0357	0.0357
Non-Wood	*Zea Mays*-Stover	0.0241	1	0.0241	0.0241	0.0344	0.0344
Wood	Alnus	0.6876	1	0.6876	0.6876	0.0113	0.0113
Wood	Bombax	0.0148	1	0.0148	0.0148	0.0369	0.0369
Wood	Euca	0.0641	1	0.0641	0.0641	0.0322	0.0322
Wood	Pine	0.0809	1	0.0809	0.0809	0.0316	0.0316

The trend line characteristics for the oxygen valuation optimized model across specific biomass species reveal a consistent pattern of perfect prediction fit ($R^2P = 1$) for all samples, while calibration performance (R^2C) varies significantly between species. Among non-wood biomass, rice husk ($R^2_C = 0.1724$) and bamboo (0.1539) demonstrated the strongest calibration performance, with moderate slope values, indicating some degree of alignment between measured and predicted oxygen content. Other non-wood species such as bagasse, Zea mays cob, shell, and stover showed low R^2_C values (0.0241–0.0472), with correspondingly shallow slopes, reflecting weak calibration trends. Their intercepts (0.0299–0.0357) remained within a narrow range, suggesting consistent but biased prediction levels. Within the wood group, Alnus once again showed the highest calibration performance (R^2_C = 0.6876), with a slope of 0.6876 and a low intercept (0.0113), indicating strong model fit and minimal bias, similar to its behavior in other elemental models. The remaining wood species—Bombax, Eucalyptus, and Pine—displayed low calibration R^2 values (0.0148–0.0809) and relatively small slopes, pointing to weaker trend representation. Although the perfect R^2 in prediction suggests a superficially ideal fit, the low calibration statistics across most species (especially in the non-wood group) indicate limited generalizability and possible overfitting. Overall, Alnus remains the most reliable species for oxygen modeling, while rice husk and bamboo show comparatively better trend characteristics among non-wood biomass.

3 Discussion

3.1 Interpretation of Finding

The findings of this study confirm that negative biomass values, though sometimes overlooked, are a critical source of error in ecological datasets and can significantly distort regression-based predictions of Global Warming Potential (GWP). The presence of negative biomass entries introduces biologically implausible data points that inflate model variance, mislead coefficient estimates, and degrade predictive performance. These results are consistent with Osborn and Overbay.[6] Its findings, which emphasized the disproportionate influence of outliers and data anomalies on regression analysis. By employing scatter plot analysis as a diagnostic tool, we were able to visually and systematically detect and remove negative biomass values before model training. This approach led to noticeable improvements in model performance metrics (e.g., reduced RMSE and increased R^2), suggesting that the removal of such anomalies enhances both the accuracy and interpretability of GWP prediction models. This aligns with recommendations by Kuhn and Johnson[9], who stressed that preprocessing and careful data filtering are often more critical than algorithm selection when building predictive models. Importantly, the use of scatter plot diagnostics offers a transparent and replicable method for preliminary anomaly detection. While traditional statistical methods can identify extreme values numerically, visual techniques provide additional insight by revealing patterns, clusters, and breakpoints that may otherwise go unnoticed [11]. This is especially valuable in environmental modeling, where the complexity and variability of ecological systems often make automated outlier detection less reliable. Moreover, the findings support the broader claim in the literature that model robustness depends heavily on data quality [8]. Even minor improvements in data preprocessing can lead to substantial gains in predictive accuracy and scientific reliability. The ecological context adds an additional layer of importance, as misestimating GWP contributions due to bad data can have implications for climate modeling, carbon accounting, and environmental policy decisions[12]. Nevertheless, the visual nature of scatter plot diagnostics has limitations. While effective in small to medium-sized datasets, manual inspection may become impractical for large-scale ecological inventories. This highlights a clear direction for future research: developing automated, reproducible anomaly detection algorithms that combine visual interpretability with statistical rigor.

3.2 Comparision of results to other

The predictive modeling of Global Warming Potential (GWP) based on elemental composition varies significantly between biomass types. In our study, non-wood species such as *Zea Mays* (shell), bagasse, and bamboo demonstrated notably poor correlations across all elemental predictors—Carbon, Hydrogen, and Oxygen—with R^2 values often below 0.05. These weak relationships suggest that the variability in chemical composition of non-wood biomass introduces noise and reduces the reliability of linear regression-based GWP prediction. Non-wood biomass contains highly variable chemical compositions—differences in cellulose, hemicellulose, lignin, ash, and extractives—which cause distinct and inconsistent NIR absorption features. This variability introduces spectral noise and disrupts the linear relationship between spectra and GWP, reducing the reliability of linear regression–based

predictions. This contrasts sharply with the performance observed in wood-based species, where elemental correlations with GWP were considerably stronger. As the elemental correlations with GWP were considerably stronger because GWP is directly influenced by elemental composition (e.g., C, H, N, O), which determines combustion emissions, making the relationship more consistent and predictable than with bulk chemical components.For example, Alnus exhibited higher R^2 values: 0.69 for Carbon, 0.57 for Hydrogen, and 0.68 for Oxygen figure 1. Such results indicate that wood species present more consistent and predictable elemental relationships, making them more suitable for inclusion in linear regression models predicting GWP. These findings diverge from the assumptions of the Yin equation, which was originally developed using data from wood-based biomass. The Yin model implicitly assumes homogeneity in elemental contributions to GWP across biomass types. The Yin model implicitly assumes that the contribution of each element (e.g., C, H, O) to GWP is consistent across all biomass types. In reality, biomass types differ in structure, combustion behavior, and trace components, so this homogeneity assumption may not hold, potentially leading to bias when applied to mixed-species datasets. However, our results suggest that this assumption does not hold for non-wood species, where the elemental variance is likely influenced by agricultural origin, processing conditions, or anatomical structure. For wood biomass, elemental variance is generally lower than in agricultural residues because trees of the same species have more uniform anatomical structure and chemical composition. However, factors like growth conditions (soil, climate), age, and specific wood part (heartwood vs. sapwood) can still influence elemental content, though typically to a smaller extent than in non-wood biomass.This mismatch may explain the poor performance of the Yin equation when applied to non-wood species. Importantly, by identifying and removing non-wood samples with low elemental correlation through scatter plot analysis, our study was able to improve model performance. The regression plots exhibited steeper and more consistent slopes, with significantly increased R^2 values for the remaining wood-based data. This not only enhanced model interpretability but also improved its predictive capability. Overall, our findings reinforce the need for species-specific or category-specific modeling approaches rather than universal equations like Yin's, especially when dealing with heterogeneous biomass datasets. Preliminary scatter plot analysis serves as a valuable diagnostic tool to identify and eliminate samples that act as confounding noise in predictive modeling results with other

3.3 Significance of Data Quality on Predictive Modeling

Data quality plays a foundational role in the success and credibility of predictive modeling. Regardless of the sophistication of the modeling technique, poor-quality data can undermine the accuracy, reliability, and interpretability of predictions. High-quality data—defined by its accuracy, completeness, consistency, and relevance—serves as the bedrock of effective model training and evaluation. As the greatest gains in model performance often come not from changing algorithms, but from improving data preprocessing and quality[9]. This is particularly true in environmental modeling contexts, where datasets are often assembled from field measurements, remote sensing, or compiled inventories—all of which are prone to measurement error, missing values, and biologically implausible entries such as negative biomass. Data anomalies such as outliers, extreme values, or coding errors can distort statistical relationships, leading to biased parameter estimates, inflated residuals, and reduced generalizability of the model [6]. For instance, in regression models, even a few high-leverage data points can exert disproportionate influence on the slope of the regression line, compromising the predictive utility of the model[13]. Moreover, the consequences of poor data quality are magnified in high-stakes applications like climate change modeling, healthcare analytics, and financial forecasting. In these domains, inaccurate predictions can result in flawed policy, misallocated resources, or environmental mismanagement [14]. As emphasized by[8], Little and Rubin (2019), careful treatment of missing and anomalous data is not just the best practice, it is essential for drawing valid conclusions. To address these challenges, data cleaning, outlier detection, and imputation techniques should be seen as integral components of the modeling pipeline, not as peripheral tasks. The growing use of visual diagnostics, such as scatter plot analysis, and automated anomaly detection methods highlights the continuing importance of data quality assurance in both traditional statistical and modern machine learning workflows[15].

3.4 Scatter Plot Analysis as a Low-Cost, High-Impact Diagnostic Tool

Scatter plot analysis remains one of the most accessible yet powerful tools in the data analyst's toolbox. Despite its simplicity, it provides immediate visual insight into relationships, anomalies, trends, and data integrity that are not always evident through statistical summaries alone. In the context of predictive modeling, particularly regression-based approaches, scatter plots offer a low-cost, high-impact method for detecting structural issues in datasets, such as outliers, non-linearity, and biologically implausible values like negative biomass. Cleveland, a pioneer in data visualization, emphasized that graphical methods such as scatter plots enable analysts to detect patterns and deviations that can guide both model selection and data cleaning strategies. These visual diagnostics help determine whether a linear model is appropriate, identify clusters or subgroups within the data, and flag unusual observations that could bias model outcomes. In environmental data modeling, where datasets often originate from field measurements or compiled databases, scatter plots serve as a crucial quality-control function. They allow researchers to visually inspect the validity of ecological relationships—such as the expected positive correlation between biomass and carbon sequestration—and to

identify data entry or measurement errors early in the analysis pipeline. Moreover, scatter plots enhance transparency in the modeling process. Their interpretability makes them valuable not only for technical analysts but also for interdisciplinary teams, policymakers, and stakeholders who require visual validation of model assumptions and findings. As such, they embody the principle of "exploratory data analysis" introduced by Tukey (1977), which prioritizes visual and interactive data exploration as a precursor to formal modeling. Although more advanced anomaly detection methods exist, scatter plot analysis offers a computationally inexpensive, universally applicable, and highly interpretable method for pre-modeling diagnostics. When integrated early in the workflow, it can prevent costly errors and improve model robustness with minimal resource investment.

3.5 Limitations of Scatter Plot-Based Data Cleaning

While scatter plot analysis offers a highly accessible and interpretable method for diagnosing data issues, it is not without limitations. One major concern is the subjectivity inherent in visual interpretation. Analysts may differ in their judgment of what constitutes an outlier or anomalous pattern, leading to inconsistencies in how data points are classified and treated. As Unwin, Theus, and Hofmann (2006) observe, even experienced users can introduce bias when interpreting plots, particularly in complex or high-dimensional datasets. Another significant limitation is the risk of data loss through manual or overly aggressive outlier removal. In ecological and environmental datasets, variability is often high, and what may appear to be an anomaly could in fact reflect real biological variation. Removing such data without proper statistical justification can lead to underfitting, reduced generalizability, and suppression of meaningful signals. Over-cleaning may also disproportionately impact minority species or rare events—precisely the types of data that are often of high ecological importance. Furthermore, scatter plots scale poorly with large datasets, where overplotting and clutter can obscure meaningful insights. While newer visualization techniques such as hexbin plots or interactive dashboards can mitigate these issues, they also require more computational resources and technical expertise, reducing the low-cost advantage that scatter plots typically offer. Lastly, the lack of automation in visual diagnostics poses a barrier to reproducibility and scalability. Unlike algorithmic outlier detection methods that can be codified and validated, visual analysis is difficult to replicate exactly across researchers or studies, which weakens methodological transparency. These limitations suggest that while scatter plot analysis is a valuable tool for initial diagnostics, it should be used in conjunction with objective statistical criteria and domain expertise to ensure rigorous and responsible data preprocessing.

The scatter plot of measured versus predicted values derived from NIR spectroscopy provides a visual evaluation of the model's predictive performance. In this study, each point on the plot represents a sample where the x-axis indicates the predicted value from the NIR model, and the y-axis denotes the corresponding reference measurement (GWP). A strong linear correlation was observed in the scatter plots in fig 2, with most data points closely aligned along the regression line. This alignment along the regression line cannot imply a high level of agreement between predicted and actual values, and not related any whether NIR spectra contain sufficient chemical information to accurately model the target property or not, though. Minimal dispersion around the line further supports the precision of the prediction.

4 Conclusion

This study demonstrates the critical role of scatter plot regression analysis in improving the accuracy of Global Warming Potential (GWP) prediction from biomass constituents. By identifying and removing biomass species with low or negligible correlation ($R^2 < 0.05$) between elemental composition and GWP—particularly among non-wood types such as Bagasse, Bamboo, and *Zea Mays* derivatives—the overall model performance significantly improved. In contrast, wood species like Alnus provided strong linear relationships between elemental composition and GWP affirming their value in predictive modeling. The removal of negative samples enhanced regression clarity, increased R^2 values, and yielded more meaningful slope coefficients, thereby improving the model's reliability and interpretability. This approach is especially valuable when developing NIR-based predictive models, where irrelevant or misleading samples can compromise chemometric accuracy. Ultimately, this study reinforces the importance of preliminary visual and statistical screening for optimizing biomass datasets in sustainable energy research.

References

1. V. Masson-Delmotte, et al., Ipcc, 2021: Summary for policymakers. in: Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change. 2021.
2. A.I. Osman, et al., Conversion of biomass to biofuels and life cycle assessment: a review. Environmental chemistry letters. **19(6)**, 4075-4118 (2021) https://doi.org/10.1007/s10311-021-01268-0
3. A. Demirbaş, Biomass resource facilities and biomass conversion processing for fuels and chemicals. Energy conversion and Management.. **42(11)**, 1357-1378 (2001) https://doi.org/10.1016/S0196-904(00)00137-0
4. F. Cherubini, et al., CO2 emissions from biomass combustion for bioenergy: atmospheric decay and contribution to global warming. Gcb Bioenergy. **3**(5), 413-426 (2011) https://doi.org/10.1111/j.1757-1707.2011.01102.x

5. A.A. Jamali, R. Ghorbani Kalkhajeh, Urban environmental and land cover change analysis using the scatter plot, kernel, and neural network methods. Arabian Journal of Geosciences. **12**(3), 100 (2019) https://doi.org/10.1007/s12517-019-4263-2
6. J.W. Osborne, A. Overbay, The power of outliers (and why researchers should always check for them). Practical Assessment, Research, and Evaluation. **9(1)**, 6 (2004) https://doi.org/10.7275/qf69-7k43
7. R.B. Kline, Principles and practice of structural equation modeling, (Guilford publications, 2023)
8. R.J. Little, D.B. Rubin, Statistical analysis with missing data, (John Wiley & Sons, 2019)
9. M. Kuhn, K. Johnson, Applied predictive modeling, (Springer New York, NY, 2013) https://doi.org/10.1007/978-1-4614-6849-3
10. B. Shrestha,et al., Comprehensive Assessment of Biomass Properties for Energy Usage Using Near-Infrared Spectroscopy and Spectral Multi-Preprocessing Techniques. Energies. **16(14)**, 5351 (2023) https://doi.org/10.3390/en16145351
11. W.S. Cleveland, Visualizing data, (Hobart press, 1993)
12. Y. Pan, et al., A large and persistent carbon sink in the world's forests. science. **333(6045)**, 988-993 (2011) https://doi.org/10.1126/science.1201609
13. Weisberg, S., Applied linear regression. Vol. 528. 2005: John Wiley & Sons.
14. Batini, C., et al., Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 2009. **41**(3): p. 1-52. https://doi.org/10.1145/1541880.1541883
15. García, S., et al., Big data preprocessing: methods and prospects. Big data analytics. **1(1)**, 9 (2016) https://doi.org/10.1186/s41044-016-0014-0