Article

Tree-based Modeling for Large-scale Management in Agriculture: Explaining Organic Matter Content in Soil

Woosik Lee and Juhwan Lee *

Posted Date: 26 January 2024

doi: 10.20944/preprints202401.1840.v1

Keywords: agricultural data analysis; agricultural business management; tree-based models; SHAP; soil organic matter; economic analysis

Article

Tree-Based Modeling for Large-Scale Management in Agriculture: Explaining Organic Matter Content in Soil

Woosik Lee [1] and Juhwan Lee [2,*]

[1] Department of Smart Distribution and Logistics, Gyeongsang National University, Jinju 52725, Republic of Korea; woosiklee@gnu.ac.kr

[2] Department of Smart Agro-Industry, Gyeongsang National University, Jinju 52725, Republic of Korea; juhwan.lee@gnu.ac.kr

* Correspondence: juhwan.lee@gnu.ac.kr; Tel.: +82-55-772-3564

Abstract: Machine learning has become more prevalent as a tool used for biogeochemical analysis in agricultural management. However, a common drawback of machine learning models is the lack of interpretability, as they are black boxes that provide little insight into agricultural management. To overcome this limitation, we compared three tree-based models (decision tree, random forest, and gradient boosting) to explain soil organic matter content through Shapley additive explanations (SHAP). Here, we used nationwide data on field crops, soil, terrain, and climate across South Korea (n = 9,584). Using the SHAP method, we identified common primary controls of the models, for example, regions with precipitation levels above 1400 mm and exchangeable potassium levels exceeding 1 $\mathrm{cmol^{+}\ kg^{-1}}$, which favor enhanced organic matter in the soil. Different models identified different impacts of nutrients on the organic matter level in the soil. The SHAP method is practical for assessing whether different machine learning models yield consistent findings in addressing these inquiries. Increasing the explainability of these models means determining essential variables related to soil organic matter management and understanding their associations for specific instances.

Keywords: agricultural data analysis; agricultural business management; tree-based models; SHAP; soil organic matter; economic analysis

1. Introduction

Machine learning (ML) has emerged as a crucial domain in science and technology, exerting a substantial socioeconomic–environmental influence on various aspects of human and natural systems [1,2]. ML allows us to learn from vast amounts of data and improve the predictive performance of models. However, ML often employs complex algorithms, which results in black-box models due to their intricate internal processes that are not readily interpretable [3–5]. Such opaqueness may lead stakeholders to overlook meaningful patterns or issues arising from hidden biases in the data. This hinders the handling of effective predictive resource management, mainly when based on large-scale ML [6]. For example, the accuracy problem of inaccurate yield mapping is well-known due to errors inherent to a high data volume and algorithms' opacity [7].

Considerable concern has been expressed about relying on opaque models that may result in decisions that are not fully comprehended or, even worse, violate ethical principles in terms of business and the environment or legal norms [1,8]. These risks are particularly relevant for critical decision making in real-life scenarios and for access to public benefits [9], for example, digitalization in agriculture [10] and terrestrial conservation [11]. This partly explains the relatively low adoption rate of current ML-based decision support systems in many areas. Land managers, government agencies, and companies that incorporate black-box ML models into their practices, products, and applications potentially face efficiency, safety, and trust issues [12]. Therefore, the lack of interpretability must be addressed, and increasing the explainability of predictive modeling for data analysis is becoming increasingly important for mitigating these unintended risks and promoting the correct application of ML models in critical domains.

In 2018, the European Parliament implemented the General Data Protection Regulation, which establish provisions regarding automated decision making [1]. These regulations aim to ensure that individuals have the right to receive "comprehensible explanations of the underlying reasoning" when automated decision-making processes are used. Additionally, in 2019, the European Union's High-Level Expert Group on Artificial Intelligence (AI) introduced ethical guidelines for trustworthy AI, and one of the critical requirements is explainability [1]. This requirement has been incorporated into the proposed EU regulation known as the AI Act [13], which establishes standardized rules for AI, thus affecting ML as a subfield of AI. Similar but nonregulatory proposals exist for AI risk management in the U.S., such as the "Identifying Outputs of Generative Adversarial Networks Act" and the "National Artificial Intelligence Initiative Act of 2020" [14]. Overall, the consensus on the importance of developing practical explanation tools is growing. Meaningful explanations are critical for describing data, testing models, identifying potential biases, addressing risks, and fostering trust and collaboration between humans and their AI assistants. However, this remains an ongoing scientific challenge [5].

An optimum model should be highly accurate and easy to interpret [2]; however, despite the rising interest in these models, achieving both interpretable and highly accurate model outputs has presented a considerable challenge [15], particularly in response to the abovementioned concerns at the management and policy scales. Consequently, the development of various explanation methods has increased for black-box models in both academia and industry [16–20]. Explainable ML emerged in the late 2010s for prediction in different systems to better explain black-box models and comprehensively address diverse aspects of the food and agriculture sector [21,22]. Explainable ML seeks to enhance the interpretability of complex algorithms while still maintaining their accuracy. Prioritizing interpretable predictions is more important than solely focusing on accurate predictions when using ML models, especially for decision making [9,23].

The objective of the study was to showcase the potential of explainable ML algorithms in analyzing large-scale data. Agricultural system data and models allow us to explore the biophysical, practical, and social aspects of food production [24,25]. However, black-box machine learning often determines false relationships between components in the system, making it unsuitable for predicting and explaining [26]. It is important to address these issues in this field because agricultural production is influenced by the management decisions of growers in response to changes in our climate and environment. We specifically focused on how different tree-based ML models can uncover novel patterns of organic matter in soil from data from Korea's field cropland, compiled on a national scale. Here, we targeted soil organic matter, as soils are integral to food production and have the potential to mitigate greenhouse gas emissions [27], while the soil pools are the most vulnerable to land degradation and climate change [17,18] and are constrained by various social, economic, and political factors [28]. This approach highlights the dominant controls of soil organic matter across fields in which its distribution interacts with the environmental state and the sociocultural matrix [29,30].

2. Materials and Methods

2.1. Data compilation and processing

We compiled environmental variables representing soil, terrain, climate, and vegetation to describe diverse field conditions in South Korea. We categorized explanatory variables into five groups: 1) soil chemical properties; 2) soil map; 3) terrain; 4) vegetation; and 5) climate (Table 1).

The National Institute of Agricultural Sciences (NAS) provides data on soil chemical properties (0–0.15 m depth) for agricultural fields [31]. Specifically, the data include organic matter (g kg^{-1}), pH (1:5 H_2O), available P_2O_5 (mg kg^{-1}), available SiO_2 (mg kg^{-1}), exchangeable magnesium (cmol$^+$ kg^{-1}), exchangeable potassium (cmol$^+$ kg^{-1}), exchangeable calcium (cmol$^+$ kg^{-1}), and electric conductivity (dS m^{-1}) [31]. The soil data are updated annually across fields to recommend a crop management plan and fertilizer application rates. The dataset is open-source and contains field-specific data for the last three years, accessible from the administrative division. Soil chemical data within each division are organized by location and sampling date. The Rural Development Administration (RDA) provides thematic soil maps of 30 physical properties at a 125 m resolution (1:25,000 scale) [32]; we selected topsoil (0–20 cm) texture, drainage class, erosion grade, soil order, soil structure, and parent material in this study.

Table 1. Description of the data inputs used as proxies for soil organic matter.

Category	Variable (abbreviation)	Unit	Resolution	Source
Soil	Organic matter (OM)	g kg^{-1}	Field	[31]
	Available phosphate (AP)	mg kg^{-1}	Field	
	Available silicate (AS)	mg kg^{-1}	Field	
	Exchangeable magnesium (Mg)	cmol$^+$ kg^{-1}	Field	
	Exchangeable potassium (K)	cmol$^+$ kg^{-1}	Field	
	Exchangeable calcium (Ca)	cmol$^+$ kg^{-1}	Field	
	pH (1:5 H_2O)		Field	
	Electric conductivity (EC)	dS m^{-1}	Field	
Soil map	Topsoil texture (TT)	class	250 m	[32]
	Drainage (DC)	class	250 m	
	Soil order (OR)	group	250 m	
	Soil structure (SS)	class	250 m	
	Parent material (PM)	Type	250 m	
	Erosion (EG)	grade	250 m	
Terrain	Elevation (DEM)	m	90 m	[33]
	Slope [1]	radians	90 m	
	Aspect [1]	radians	90 m	
	Flow direction (flowdir) [1]	m	90 m	
	Roughness [1]	m	90 m	
	Hill shade (hill) [2]		90 m	
	Topographic position index (TPI) [1]		90 m	
	Terrain ruggedness index (TRI) [1]		90 m	
	Upslope contributing area (a) [1]		90 m	
	Topographic wetness index (TWI) [1]		90 m	
Climate	Mean annual temperature (TA)	°C	1 km	[34]
	Maximum annual temperature (TAMAX)	°C	1 km	
	Minimum annual temperature (TAMIN)	°C	1 km	
	Mean annual precipitation (RN)	mm	1 km	
	Solar irradiation (SI)	MJ m^{-2}	1 km	
	Relative humidity (RHM)	%	1 km	
	Wind speed (WS)	m s^{-1}	1 km	
Vegetation	Net primary productivity (NPP)	g C m^{-2} y^{-1}	11 km	[35]

[1] Estimated based on DEM data.

[2] Computed from slope and aspect values, assuming sun elevation and direction (azimuth) angles of 45° and 0°, respectively.

We obtained the digital elevation model (DEM) data (90 m resolution), which represented a continuous topographic elevation surface, from the National Spatial Data Infrastructure (NSDI) portal [33]. We used the DEM data to derive slope, aspect, roughness, topographic position index, terrain ruggedness index, and flow direction. Hill shade was computed from slope and aspect [36], assuming sun elevation and direction (azimuth) angles of 45° and 0°, respectively. Lastly, we estimated the upslope contributing area and topographic wetness index according to Quinn *et al.* [37]. We used a global estimate of annual net primary production in TIFF format from 2019 to 2021 to represent vegetation from the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA's Terra and Aqua satellites [35]. The spatial resolution of the data is 0.1° (approximately 11 km). To represent the current climate, we used mean annual air temperature, maximum/minimum air temperature, mean annual precipitation, solar irradiation, relative humidity, and wind speed data based on the Modified Korean-Parameter-elevation Regressions on Independent Slopes Model (MK-PRISM, version 1.2). All data were available from 2000 to 2019, except for solar irradiation (2014–2019) [34]. The MK-PRISM data were in netCDF format. All climatic variables were averaged annually and then over the entire period, whereas the sum of precipitation values was calculated and averaged over the same period.

According to the land cover data from EGIS [38], agricultural land subclasses include: 1) fields (including rearranged fields); 2) rice paddies; 3) facility cultivation; 4) orchards; and 5) pastures and nurseries. Based on the survey of arable land in 2021 [39], the total field area was 766,000 ha. We first compiled valid soil chemical data and the coordinates of the fields (n = 263,790). We then extracted the values of the climate, terrain, and soil map variables with the coordinates of the samples. We removed outliers from the complete dataset based on Cook's distance (n = 9,584) and scaled the data for modeling. All data were compiled and processed in R [40].

2.2. Explainable Tree-based Models

We chose decision tree (DT), random forest (RF), and gradient boosting (GB) due to their ease of interpretation and proven success in handling structured datasets [41,42]. The DT algorithm is a recursive partitioning method that iteratively generates child nodes, which are further divided into pairs of nodes. DT serves as an interpretable ML model, as the decision path from the root to the leaf or terminal nodes keeps track of the features used for predictions. However, DT learning is difficult to comprehend and suffers from overfitting when the models become more complex (i.e., with a larger maximum depth). Ensemble methods, such as RF and GB, are used instead to address this issue. By incorporating multiple decision trees, these approaches produce robust and accurate models. Leveraging the diversity of individual trees, they mitigate overfitting while enhancing overall performance. RF uses feature randomization to create each tree. At each split, a random subset of features is considered to determine the best splitting point for the node. This randomness increases the diversity among trees and enhances the generalization capability of the ensemble model.

GB is an alternative ensemble learning technique in which an additive model is constructed by sequentially combining predictions from multiple decision trees, thus creating a robust predictive model. RF primarily aims to reduce variance through subset (i.e., bagging) and feature randomization. In contrast, GB focuses on diminishing bias and enhancing predictive accuracy through iterative optimization processes. Among the models, the DT algorithm is relatively simple and its interpretability is high. The last two models are composed of many trees that have been combined, which, therefore, must be supplemented with interpretable methods for understanding model behavior.

2.2.1. Model evaluation

We integrated the Shapley additive explanation (SHAP) method with the optimized ML to understand the prediction criteria of the models and the level of contribution of the individual features that strongly correlate in organic matter prediction. Feature importance provides a crucial reference for feature selection and model interpretability [43]. A higher feature importance value, compared with another, implies the greater importance of the feature for generating a prediction. The SHAP method, developed by Lundberg and Lee [44], is a method for interpreting the predictions of ML models. Shapley values quantify the contribution of each explanatory variable in each instance. The term "additive" signifies that the corresponding Shapley value for each explanatory variable of an instance can be additively combined [45]. This approach provides quantitative information to explain how individual explanatory variables either positively or negatively impact the target variable of interest in the model. The following formula represents the Shapley value:

$$\boldsymbol{g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_{i}},$$

where ϕ_i is the feature attribution value ($\phi_i \in \mathbb{R}$), and z'_i represents the binary variables ($z'_i \in \{0,1\}^M$). When an explanatory variable is included and has a value of $z'_i = 1$, it becomes an excluded input feature if its value is $z'_i = 0$. M represents the total number of input features, while g explains the model. Furthermore, the importance of feature *i* is defined by the Shapley value:

$$\boldsymbol{\phi_i = \frac{1}{|N|!} \sum_{S \subseteq |N| \backslash \{i\}} |S|\,!\,(|N| - |S| - 1)!\,[f(S \cup \{i\}) - f(S)]},$$

where f(*S*) represents the output of the ML model being explained with a set *S* of features, and *N* refers to the entire set of available features. The Shapley value or contribution of a specific feature *i*(ϕ_i) is calculated as the average contribution across all possible permutations of the feature set [46].

Although the Shapley-based feature importance method effectively identifies important features, it does not provide insights into two crucial aspects: 1) how changes in these important features affect predictions and 2) whether a specific threshold for these key features can enhance the accuracy of predicting the target variable. To address these concerns, we employed partial dependence (PD) plotting, which enabled us to analyze how specific input features contribute to variations in the expected response of the target variable. One-dimensional PD plots illustrate how changes in the chosen independent variables impact the expected value of their dependent variable while other independent variables remain constant. Furthermore, two-dimensional PD plots are used to assess predicted values for their dependent variable as both independent variables simultaneously vary.

PD plots are a tool for comprehending the impacts of features in any machine learning model [47], such as $\hat{f}$(x). The explanatory variables $x = (x_s,\ x_c)$ can be divided into two subsets to emphasize the concept and effectively convey this relationship and visualization: x_s and x_c. x_s represents the "chosen" independent variable, and x_c comprises the set of other independent variables. The function that represents PD is defined as:

$$PD_S(\mathrm{X}) = PD_S(X_S) = \int \hat{f}(X_S, X_C) dP(X_C).$$

The PD curve is estimated by calculating the average prediction when a specific variable subset, denoted as X_S, takes on specific values, represented as x_S [47].

$$\widehat{PD}_S(x) = \widehat{PD}_S(x_s) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_s, x_{ic}).$$

Note that, when X_S is equal to *X*, the value of the partial dependence function PD_S(x) corresponds to the estimated output function $\hat{f}$(x). In a hypothetical scenario where s is an empty set, such as when $X_C = X$, this would result in:

$$PD_\emptyset(X) = PD_\emptyset = \int \hat{f}(X) dP(X),$$

which does not depend on x and represents the constant average prediction of the model calculated as:

$$\widehat{PD}_\emptyset(x) = \widehat{PD}_\emptyset = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_i).$$

2.2.2. Assessment statistics

To train the model, we performed five-times repeated five-fold cross validation. Grid search was employed for exhaustive searching to systematically tune the hyperparameters in ML from a predefined set of values to enhance performance and mitigate the risk of overfitting (Table 2).

Table 2. Optimal hyperparameters of three tree-based models.

Model parameter	Parameter grid	Decision Tree	Random Forest	Gradient Boosting
Maximum depth of a tree	[6,8,10,12]	6	12	8
Minimum samples per leaf	[8,12,18]	12	8	18
Minimum number of samples	[8,16,20]	8	8	8
Number of trees	[10,100]	-	100	100

To assess the performance of the models, we computed the root mean squared error (RMSE) to quantify the inaccuracy of the estimates, the mean absolute error (MAE) to quantify the errors between modeled and observed values, and the coefficient of determination (R^2). The RMSE accounts for both the bias and the imprecision of the analysis. We report the mean and standard deviation of the assessment statistics from the cross validations.

3. Results

3.1. Comparison of Selected Tree-based Models

Among the models, GB exhibited the superior performance, as indicated by the highest R^2 of 0.59 (Figure 1). Following closely, RF demonstrated reasonable predictive ability with an R^2 of 0.57. Lastly, the DT model displayed the weakest performance, with an R^2 of 0.45. Furthermore, the RF and GB models exhibited higher prediction accuracy with fewer errors. Moreover, these two models showed a more concentrated distribution of estimates around the observations, suggesting increased consistency and reliability in their predictions compared with those of the DT model (Figure 1).

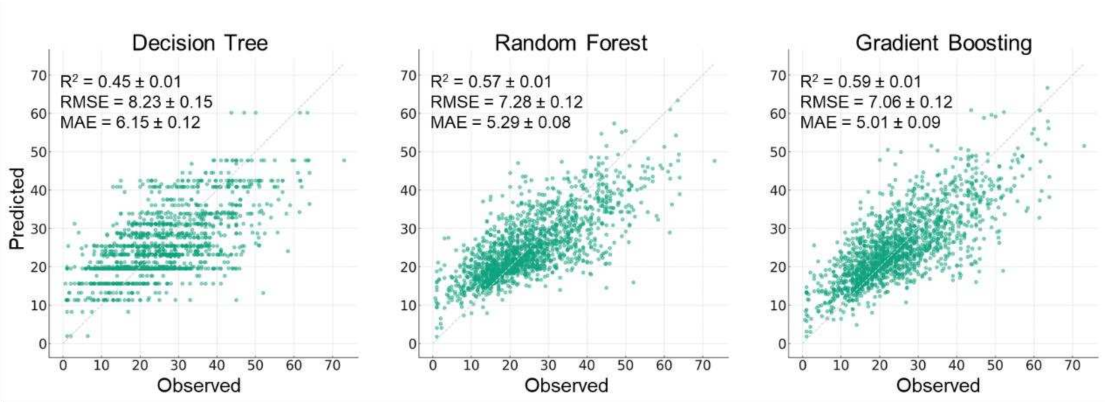

Figure 1. Scatter plots of predicted-versus-observed organic matter stocks in soil by using tree-based model. Points are field-wide averages for last three years (2020–early 2023).

The optimal values for the hyperparameters of each model used in the study are summarized in Table 2. For the DT model, a maximum depth of six was selected to prevent overfitting caused by excessive expansion of the tree structure. The minimum number of samples per node was set to 12. In the RF model, a maximum depth of 12 and a minimum number of eight samples per leaf were chosen. Additionally, an internal node requires at least eight samples for splitting to manage complexity and minimize overfitting concerns. Our RF and GB models comprised 100 trees to ensure accurate results. However, adding more trees beyond this point yielded no further improvements.

3.2. Feature Overall Importance Analysis

The global feature explanation obtained from the SHAP method for the three tree-based models is shown in Figures 2 and 3. We represent the average impact of each variable on soil organic matter based on the global feature importance, which is the mean absolute representative SHAP value for that feature over all the given samples (Figure 2). Figure 3 illustrates the impact of the features on the model's predictions. Higher SHAP values for precipitation, indicated in red, signify a large and positive average impact on predictions, suggesting a higher likelihood of increasing the prediction accuracy due to the high mean absolute SHAP values.

The features are ordered from top to bottom by their predictive importance. In this case, we found that all models consistently identified the mean annual precipitation (RN) as the primary predictor. This was closely trailed by exchangeable K content in the DT and RF models (Figure 3). Similarly, the GB model acknowledged exchangeable K as one of the top five predictors; however, it attributed relatively less importance to it (Figure 3).

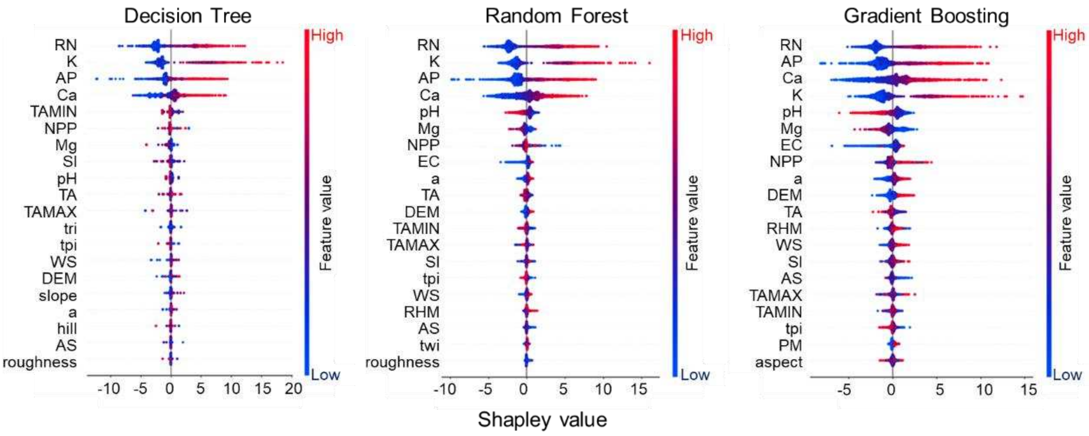

Figure 2. Global feature explanation through Shapley additive explanation for three tree-based models. The range of Shapley values was computed for each explanatory variable to determine its impact on soil organic matter content. For abbreviations, refer to Table 1.

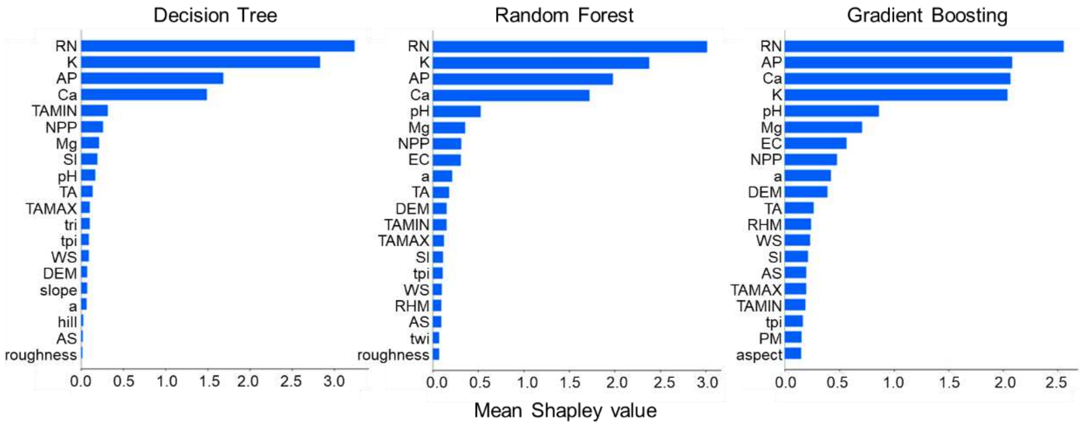

Figure 3. Shapley-based feature importance of environmental variables for global explanations of soil organic matter content, expressed as mean Shapley values. For abbreviations, refer to Table 1.

3.2. Feature Partial Dependence Analysis

Subsequently, we undertook a more in-depth investigation into the impacts of precipitation and exchangeable K on soil organic matter content. This choice was driven by the results of the DT and RF analyses, which highlighted these variables as being within the top two regarding Shapley-based feature importance (Figures 2 and 3). In Figure 4, we present PD plots illustrating how the three models portrayed the connections between precipitation and organic matter content; all models consistently showed a positive correlation between them. Nonetheless, we identified disparities in the strength and form of the relationships across the models. Specifically, the DT model exhibited a

doi:10.20944/preprints202401.1840.v1

distinct, step-like association centered around a precipitation value of 1400 mm (Figure 4). However, both RF and GB models suggested a positive yet nonlinear relationship.

When scrutinizing the association between soil organic matter content and other variables, all three models concurred on the positive nature of these relationships for precipitation, K, and available P. Notably, only the DT model revealed a conspicuous stepwise linkage, discernible at the values of precipitation (1400 mm), K (1 cmol$^+$ kg^{-1}), and available P (600 mg kg^{-1}).

Two-dimensional PD plots were employed to visually determine the interaction effects between precipitation and exchangeable K (Figure 5). All models indicated that organic matter is contingent upon both precipitation and K. The observed patterns across all models displayed distinct divisions in the patterns along the K contents around the value of 1 cmol$^+$ kg^{-1} and along the amount of precipitation of approximately 1400 mm. This observation strongly suggests the presence of an interaction effect between these variables. The presentation of PD plots focusing on precipitation while conditioning on varying K thresholds effectively demonstrates this interaction effect in Figure 5. A discernible pattern emerges from these plots, showing that the association with precipitation exhibits heightened strength when the level of exchangeable K surpasses 1 cmol$^+$ kg^{-1}.

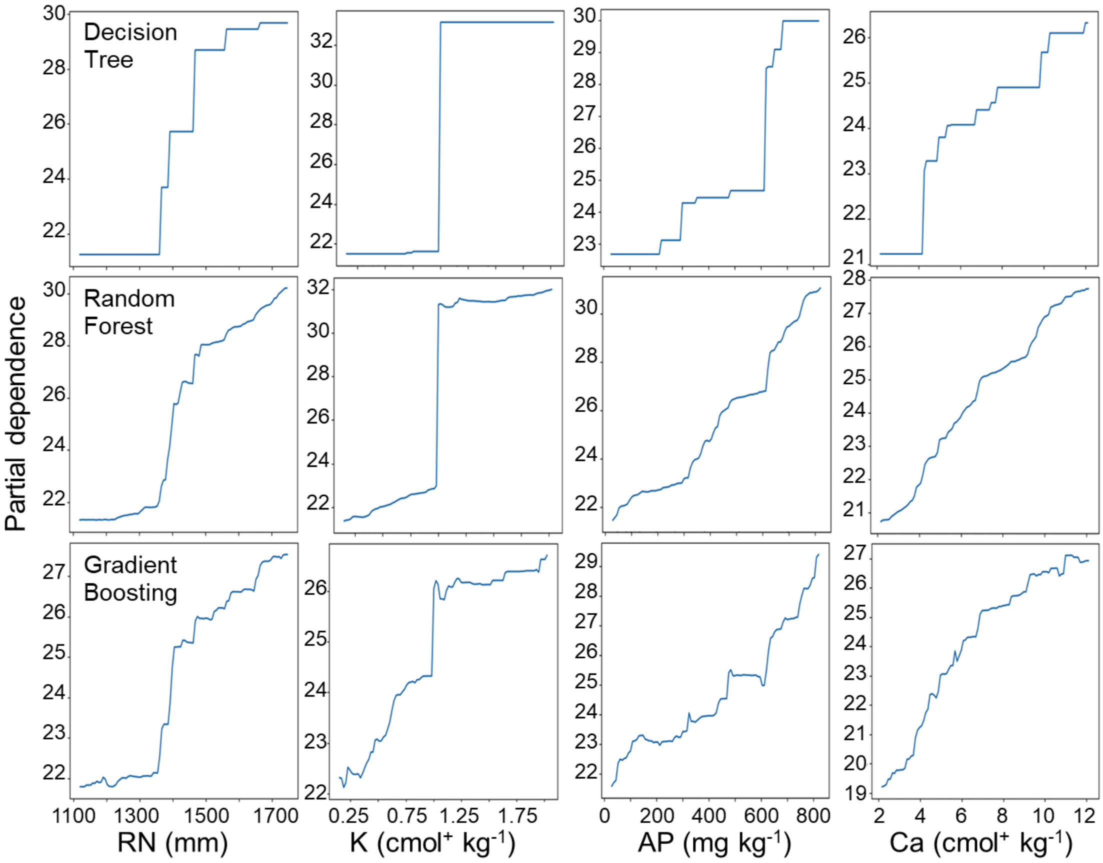

Figure 4. Partial dependence plots of soil organic matter content with precipitation (RN), exchangeable potassium (K), available P_2O_5 (AP), and exchangeable calcium (Ca) (global explanations).

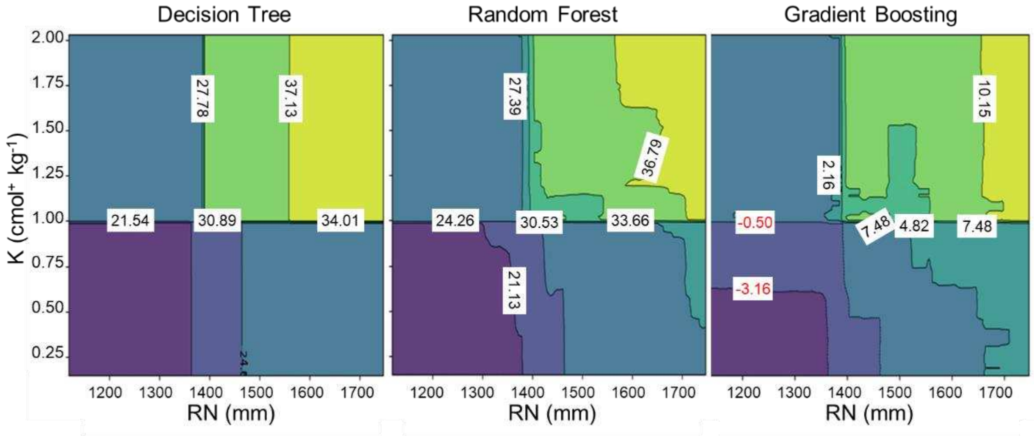

Figure 5. Partial dependence of soil organic matter between precipitation (RN) and exchangeable potassium (K).

4. Discussion

Based on the results of the model performance analysis, we found that GB yielded the highest accuracy among the models evaluated, followed by RF and then DT. The precipitation variable (RN) consistently emerged as the most influential predictor in all these models under Korean conditions. Subsequently, our analysis showed that soil organic matter levels under field crops also depend on soil fertility properties, such as available K, available P, and Ca levels. In addition, analyzing the bee swarm summary plots provided a more comprehensive understanding of how other variables influenced the prediction (Figure 2). All models showed positive relationships between nutrients and soil organic matter, but the intensity and shape of these relationships varied among the models. For example, the two-dimensional PD plots illustrated the divergent interaction effects of precipitation and exchangeable K (Figure 5). The findings suggested that the relationship with organic matter is more robust when exchangeable K exceeds 1 cmol⁺ kg^{-1} and annual precipitation surpasses 1400 mm. Thus, organic matter content is contingent upon both precipitation and K on a large scale. This assessment provides a comprehensive understanding of the varying importance levels attributed to key predictors using distinct modeling approaches. However, the impact of these features on the model is still not fully understood [48].

The SHAP method was also employed to assess the importance of variables at a specific local site (data not shown). The results indicated that pH was more important than RN, identified as the most influential factor in controlling global model behavior. Only the DT model revealed a conspicuous stepwise linkage, discernible at certain values of precipitation (1400 mm), available P (600 mg kg^{-1}), and K (1 $cmol^{+}$ kg^{-1}) (Figure 4). This PD analysis underscores the nuances in how different models represent the intricate interplay between soil variables and organic matter. This insight provides valuable clarity regarding these two variables' complex interplay and collective impact on soil organic matter. Up to this point, our focus was on elucidating the overarching behavior of global models, aiming to comprehend the insights derived from the data. However, this approach falls short in explaining the nuances of local model behavior, a critical aspect in discerning the factors deemed important by the models when predicting values for specific instances.

Tree-based modeling can lead to the reliable prediction of organic matter content in soil and the identification of vital environmental factors that affect organic matter content. This knowledge is essential for managing the soil’s health and the sustainability of cropping systems, and for fine-tuning fertilizer and water use management. Moreover, when addressing agricultural challenges, especially with adopting alternative production systems [49,50], the results can support the selection of a farm location to improve not only soil organic matter and fertility conditions but also the marketability and profitability of crop harvests. For example, in Kenya, agricultural practices include the use of fertilizers, pesticides, and irrigation to enhance soil organic matter, potentially producing premium-market-priced organic products [51]. When premium prices are available for organic produce, the organic system yields significantly higher net returns than conventionally managed systems, achieving a gross margin that is 1.3 to 4.1 times higher. Additionally, intercropping various crops enhanced overall productivity and profitability [51]. Thus, these results provide insights into the economic implications of choosing an alternative farming system based on soil organic matter levels and related conditions. In enhancing soil organic matter content, several viable management approaches may offer additional means to boosting agricultural productivity, including soil amendment; yet, the economic evaluation of each of these methods has been limited, and a gap exists in terms of the comprehensive evaluation of the socioeconomic impacts, necessitating further research in this area. Petersen and Hoyle [52] modeled the benefits of soil organic carbon, mainly focusing on the increased availability of nitrogen and increased plant-available water-holding capacity (PAWC). The value of soil organic carbon is estimated to be between AUD 7.1 and 8.7 Mg^{-1} ha^{-1} annually [134]. This valuation includes approximately 75% for carbon sequestration and smaller proportions for productivity improvements. The enhancements in PAWC (~5%) and nitrogen replacement value (~20%) contribute to higher agricultural productivity. An increased PAWC allows for increased water retention in the soil, and increased nitrogen availability supports healthier plant growth. Both soil quality and profit result in higher-quality land having a higher market value, providing an additional incentive to adopt a soil-conserving crop production system.

Mikhailova *et al.* [53] evaluated the monetary value of soil organic carbon stocks in the U.S., considering various factors such as soil order, depth, and geographic region. They estimated the total value of soil organic carbon storage to range from USD 4.64 trillion to USD 23.1 trillion. This valuation highlights the critical role of its management in delivering environmental and economic benefits. Similarly, Dube *et al.* [54] offered a detailed financial analysis of ecosystem services from healthy soils in Vermont, highlighting benefits such as increased carbon storage (USD 19 acre^{-1} year^{-1} in climate mitigation), reduced phosphorus losses (USD 8 acre^{-1} year^{-1} in water quality), erosion control (USD 2 acre^{-1} year^{-1} in waterway damage reduction), and enhanced water retention (USD 2 acre^{-1} year^{-1} in flood damage reduction), cumulatively valued at over USD 25 million annually. This emphasizes the economic importance of soil health investment and preservation. Hacisalihoglu *et al.* [55] used the "market value of soil" method to calculate the cost of soil erosion, considering nutrient loss and fertilizer market prices. They estimated an average economic loss of USD 59.54 per hectare per year in pasture lands and USD 102.36 in agricultural lands in Turkey due to soil erosion. This erosion, which tends to remove the nutrient- and organic-matter-rich topsoil, diminishes soil fertility, provides nutrients, sustains structure and moisture, and affects economic values by depleting a crucial soil component. Additionally, Kane *et al.* [56] found that counties with higher soil organic matter levels had increased yields, lower yield losses, and lower crop insurance payout rates. A 1% increase in soil organic matter corresponded to a yield boost of 2.2 ± 0.33 Mg ha^{-1} and a notable reduction of 36 ± 4.76% in the average proportion of liabilities paid. Sparling *et al.* [57] quantified the monetary value of soil organic matter in enhancing crop production in New Zealand soils. This value was determined by estimating the worth of dairy milk solids, derived from a computer simulation modeling the yield of dry pasture matter and the accumulation of organic matter. The findings revealed that soils with lower organic matter levels yielded between 8.5 to 47.7 kg fewer milk solids per hectare annually, translating to a financial impact of NZD 27 to NZD 150 per hectare. Over recovery periods of 36, 90, and 125 years, the cumulative loss per hectare at Pukekohe, due to reduced productivity, was estimated at NZD 1239 with a 3.5% discount factor and NZD 772 with a 10% discount.

Finally, Fan *et al.* [58] showed that field practices with varying organic matter inputs could affect the total ecosystem service valuation in organic cereal crop production systems. This impact likely stems from altered soil properties due to long-term diverse field management. The authors estimated the economic value of ES in these systems under different management strategies, finding that the economic values ranged from USD 1492 to USD 1969 per hectare per year. Reyes and Elias [59] demonstrated that drought and excess precipitation were the primary causes of crop losses in the U.S. from 2001 to 2016, leading to over USD 440 billion in economic damage. These studies emphasize the importance of understanding and predicting environmental factors in agriculture. Therefore, tree-based modeling for predicting organic matter in soil and identifying key factors can help improve soil health and contribute to more sustainable and profitable agricultural practices.

5. Conclusions

We addressed the lack of interpretability of ML modeling to gain more insight into agricultural management. Specifically, we selected three tree-based models (decision tree, random forest, and gradient boosting) and focused on soil organic matter, which can provide multiple benefits as a valuable resource but represents challenges commonly faced in agriculture. We found that tree-based explainable ML enables the reliable prediction of the organic matter content in soil and identifies vital environmental factors that are relevant to organic matter management. This knowledge is essential for managing soil health and the sustainability of cropping systems and for fine-tuning fertilizer and water use management. We further examined the potential interactions between selected variables within the top four as determined by the Shapley-based feature importance. Based on our results, incorporating deep learning into our analytical framework would provide considerable benefit. Deep learning has demonstrated remarkable capabilities in handling complex data patterns and may provide enhanced insights when applied to agricultural science. Our prior emphasis on overall feature importance analysis should complement the value of local study. Whereas overall feature importance analysis provides valuable insights at the global level, local analysis allows for a more granular understanding of how specific features impact the system under investigation. This perspective can lead to more precise and actionable recommendations. Lastly, employing feature elimination represents a meaningful avenue for analysis. This approach can aid with dimensionality reduction and identifying critical features, potentially simplifying the model while preserving its predictive power. Such streamlining can enhance model interpretability and efficiency, which are paramount in agricultural science.

Funding: This study was funded by the National Research Foundation of Korea (NRF), NRF-2022R1F1A1073973.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during and analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Bodria, F.; Giannotti, F.; Guidotti, R.; Naretto, F.; Pedreschi, D.; Rinzivillo, S. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery* **2021**, *37*, 1719 - 1778.
2. Ryo, M. Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artificial Intelligence in Agriculture* **2022**, *6*, 257-265, doi:10.1016/j.aiia.2022.11.003.
3. Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artif. Intell.* **2017**, *267*, 1-38.
4. Belle, V.; Papantonis, I. Principles and practice of explainable machine learning. *Frontiers in Big Data* **2021**, *4*, doi:10.3389/fdata.2021.688969.
5. Saeed, W.; Omlin, C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems* **2023**, *263*, 110273, doi:10.1016/j.knosys.2023.110273.
6. Wang, M.; Fu, W.J.; He, X.N.; Hao, S.J.; Wu, X.D. A survey on large-scale machine learning. *Ieee T Knowl Data En* **2022**, *34*, 2574-2594, doi:10.1109/Tkde.2020.3015777.
7. Visser, O.; Sippel, S.R.; Thiemann, L. Imprecision farming? Examining the (in)accuracy and risks of digital agriculture. *Journal of Rural Studies* **2021**, *86*, 623-632, doi:10.1016/j.jrurstud.2021.07.024.
8. Dundon, S.J. Agricultural ethics and multifunctionality are unavoidable. *Plant Physiol* **2003**, *133*, 427-437, doi:10.1104/pp.103.029124.

9. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **2019**, *1*, 206-215, doi:10.1038/s42256-019-0048-x.
10. Finger, R. Digital innovations for sustainable and resilient agricultural systems. *Eur Rev Agric Econ* **2023**, *50*, 1277-1309, doi:10.1093/erae/jbad021.
11. Hoang, N.T.; Taherzadeh, O.; Ohashi, H.; Yonekura, Y.; Nishijima, S.; Yamabe, M.; Matsui, T.; Matsuda, H.; Moran, D.; Kanemoto, K. Mapping potential conflicts between global agriculture and terrestrial conservation. *Proceedings of the National Academy of Sciences* **2023**, *120*, e2208376120, doi:10.1073/pnas.2208376120.
12. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* **2016**, *5 2*, 153-163.
13. Schuett, J. Risk management in the Artificial Intelligence Act. *European Journal of Risk Regulation* **2023**, 1-19, doi:10.1017/err.2023.1.
14. Thomson Reuters. *LAWnB IP Exclusive Report: 2023 Domestic and International AI Regulatory and Policy Trends*; 10/24/2023 2023; pp. 5-8.
15. Breiman, L. Statistical modeling: the two cultures. *Statistical Science* **2001**, *16*, 199-231, 133, doi:10.1214/ss/1009213726.
16. Guidotti, R.; Monreale, A.; Turini, F.; Pedreschi, D.; Giannotti, F. A survey of methods for explaining black box models. *ACM Computing Surveys* **2018**, *51*, 1-42.
17. Adadi, A.; Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138-52160.
18. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **2020**, *58*, 82-115, doi:10.1016/j.inffus.2019.12.012.
19. Theissler, A.; Spinnato, F.; Schlegel, U.; Guidotti, R. Explainable AI for time series classification: A review, taxonomy and research directions. *IEEE Access* **2022**, *10*, 100700-100724, doi:10.1109/ACCESS.2022.3207765.
20. Yuan, H.; Yu, H.; Gui, S.; Ji, S. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *45*, 5782-5799, doi:10.1109/TPAMI.2022.3204236.
21. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **2019**, *267*, 1-38, doi:10.1016/j.artint.2018.07.007.
22. Pichler, M.; Hartig, F. Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution* **2023**, *14*, 994-1016, doi:10.1111/2041-210X.14061.
23. Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys* **2022**, *16*, 1-85, 85.
24. Antle, J.M.; Basso, B.; Conant, R.T.; Godfray, H.C.J.; Jones, J.W.; Herrero, M.; Howitt, R.E.; Keating, B.A.; Munoz-Carpena, R.; Rosenzweig, C.; et al. Towards a new generation of agricultural system data, models and knowledge products: Design and improvement. *Agric Syst* **2017**, *155*, 255-268, doi:10.1016/j.agsy.2016.10.002.
25. Smith, P.; Davies, C.A.; Ogle, S.; Zanchi, G.; Bellarby, J.; Bird, N.; Boddey, R.M.; McNamara, N.P.; Powlson, D.; Cowie, A.; et al. Towards an integrated global framework to assess the impacts of land use and management change on soil carbon: current capability and future vision. *Global Change Biol* **2012**, *18*, 2089-2101, doi:10.1111/j.1365-2486.2012.02689.x.
26. Hu, T.; Zhang, X.; Bohrer, G.; Liu, Y.; Zhou, Y.; Martin, J.; Li, Y.; Zhao, K. Crop yield prediction via explainable AI and interpretable machine learning: Dangers of black box models for evaluating climate change impacts on crop yield. *Agricultural and Forest Meteorology* **2023**, *336*, 109458, doi:10.1016/j.agrformet.2023.109458.
27. Paustian, K.; Lehmann, J.; Ogle, S.; Reay, D.; Robertson, G.P.; Smith, P. Climate-smart soils. *Nature* **2016**, *532*, 49-57, doi:10.1038/nature17174.
28. Lal, R. Challenges and opportunities in soil organic matter research. *Eur J Soil Sci* **2009**, *60*, 158-169, doi:10.1111/j.1365-2389.2008.01114.x.
29. Conway, G.R. The properties of agroecosystems. *Agr. Syst.* **1987**, *24*, 95-117, doi:10.1016/0308-521X(87)90056-4.
30. Spencer, J.E.; Stewart, N.R. The nature of agricultural systems. *Annals of the Association of American Geographers* **1973**, *63*, 529-544.
31. NAS. Chemical Data for Soil Test. National Institute of Agricultural Sciences, Rural Development Administration **2023**, www.data.go.kr/data/15073569/openapi.do (accessed on 01/30/2023).

32. RDA. Precision soil maps. Rural Development Administration **2023**, https://soil.rda.go.kr (accessed on 01/19/2023).
33. NSDI. Degital Elevation Model (registered on: 08/11/2020). National Spatial Data Infrastructure **2020**, http://data.nsdi.go.kr/dataset/20001 (accessed on
34. KMA. Climate Change Scenarios. **2023**, http://www.climate.go.kr/home/CCS/contents_2021/35_download.php (accessed on 12-20-2023).
35. NEO. Net Primary Productivity (1 year - TERRA/MODIS). NASA Earth Observations **2023**, https://neo.gsfc.nasa.gov (accessed on 02/20/2023).
36. Horn, B.K.P. Hill shading and the reflectance map. *Proceedings of the IEEE* **1981**, *69*, 14-47, doi:10.1109/PROC.1981.11918.
37. Quinn, P.F.; Beven, K.J.; Lamb, R. The in(a/tan/β) index: How to calculate it and how to use it within the topmodel framework. *Hydrological Processes* **1995**, *9*, 161-182, doi:10.1002/hyp.3360090204.
38. EGIS. Land Cover Maps. Environmental Geographic Information Service **2023**, https://egis.me.go.kr/intro/land.do (accessed on 01/13/2023).
39. Statistics Korea. Arable land in Korea. *Korean Statistical Information Service* **2021**.
40. R Core Team *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria, 2023.
41. Hastie, T.; Tibshirani, R.; Friedman, J. Unsupervised Learning. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Hastie, T., Tibshirani, R., Friedman, J., Eds.; Springer New York: New York, NY, 2009; pp. 485-585.
42. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016; pp. 785–794.
43. Zhang, W.; Wu, C.; Zhong, H.; Li, Y.; Wang, L. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers* **2021**, *12*, 469-477, doi:10.1016/j.gsf.2020.03.007.
44. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017; pp. 4768–4777.
45. Liu, J.; Li, C.; Ouyang, P.; Liu, J.; Wu, C. Interpreting the prediction results of the tree-based gradient boosting models for financial distress prediction with an explainable machine learning approach. *Journal of Forecasting* **2023**, *42*, 1112-1137, doi:10.1002/for.2931.
46. Rodríguez-Pérez, R.; Bajorath, J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design* **2020**, *34*, 1013-1026, doi:10.1007/s10822-020-00314-0.
47. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **2001**, *29*, 1189-1232.
48. Czerwinska, U. Interpretability of Machine Learning Models. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, Egger, R., Ed.; Springer International Publishing: Cham, 2022; pp. 275-303.
49. Pittelkow, C.M.; Liang, X.Q.; Linquist, B.A.; van Groenigen, K.J.; Lee, J.; Lundy, M.E.; van Gestel, N.; Six, J.; Venterea, R.T.; van Kessel, C. Productivity limits and potentials of the principles of conservation agriculture. *Nature* **2015**, *517*, 365-U482, doi:10.1038/nature13809.
50. Petersen, B.; Snapp, S. What is sustainable intensification? Views from experts. *Land Use Policy* **2015**, *46*, 1-10, doi:10.1016/j.landusepol.2015.02.002.
51. Adamtey, N.; Musyoka, M.W.; Zundel, C.; Cobo, J.G.; Karanja, E.; Fiaboe, K.K.M.; Muriuki, A.; Mucheru-Muna, M.; Vanlauwe, B.; Berset, E.; et al. Productivity, profitability and partial nutrient balance in maize-based conventional and organic farming systems in Kenya. *Agriculture, Ecosystems & Environment* **2016**, *235*, 61-79, doi:10.1016/j.agee.2016.10.001.
52. Petersen, E.H.; Hoyle, F.C. Estimating the economic value of soil organic carbon for grains cropping systems in Western Australia. *Soil Research* **2016**, *54*, 383-396, doi:10.1071/SR15101.
53. Mikhailova, E.A.; Groshans, G.R.; Post, C.J.; Schlautman, M.A.; Post, G.C. Valuation of soil organic carbon stocks in the contiguous United States based on the avoided social cost of carbon emissions. *Resources* **2019**, *8*, 153, doi:10.3390/resources8030153.
54. Dube, B.; White, A.; Ricketts, T.; Darby, H. *Valuation of soil health ecosystem services*; The University of Vermont: 2022.
55. Hacisalihoglu, S.; Toksoy, D.; Kalca, A. Economic valuation of soil erosion in a semi and area in Turkey. *African Journal of Agricultural Research* **2010**, *5*, 1-6, doi:10.5897/AJAR09.595.

56. Kane, D.A.; Bradford, M.A.; Fuller, E.; Oldfield, E.E.; Wood, S.A. Soil organic matter protects US maize yields and lowers crop insurance payouts under drought. *Environmental Research Letters* **2021**, *16*, 044018, doi:10.1088/1748-9326/abe492.
57. Sparling, G.P.; Wheeler, D.; Vesely, E.T.; Schipper, L.A. What is soil organic matter worth? *Journal of Environmental Quality* **2006**, *35*, 548-557, doi:10.2134/jeq2005.0230.
58. Fan, F.; Henriksen, C.B.; Porter, J. Valuation of ecosystem services in organic cereal crop production systems with different management practices in relation to organic matter input. *Ecosystem Services* **2016**, *22*, 117-127, doi:10.1016/j.ecoser.2016.10.007.
59. Reyes, J.J.; Elias, E. Spatio-temporal variation of crop loss in the United States from 2001 to 2016. *Environmental Research Letters* **2019**, *14*, 074017, doi:10.1088/1748-9326/ab1ac9.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.