*sustainability*

MDPI

*Article*

# Comparing Machine Learning Approaches for Predicting Spatially Explicit Life Cycle Global Warming and Eutrophication Impacts from Corn Production

**Xiaobo Xue Romeiko [1,*], Zhijian Guo [2], Yulei Pang [3], Eun Kyung Lee [1] and Xuesong Zhang [4]**

[1] Department of Environmental Health Sciences, University at Albany, State University of New York, One University Place, George Education Center, Rensselaer, NY 12144, USA; elee4@albany.edu
[2] Department of Mathematics, University at Albany, State University of New York, Albany, NY 12222, USA; zguo@albany.edu
[3] Department of Mathematics, Southern Connecticut State University, 501 Crescent Street, New Haven, CT 06515, USA; pangy1@southernct.edu
[4] Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD 20740, USA; Xuesong.Zhang@pnnl.gov
* Correspondence: xxue@albany.edu

Received: 30 November; Accepted: 10 February 2020; Published: 17 February 2020


**Abstract:** Agriculture ranks as one of the top contributors to global warming and nutrient pollution. Quantifying life cycle environmental impacts from agricultural production serves as a scientific foundation for forming effective remediation strategies. However, methods capable of accurately and efficiently calculating spatially explicit life cycle global warming (GW) and eutrophication (EU) impacts at the county scale over a geographic region are lacking. The objective of this study was to determine the most efficient and accurate model for estimating spatially explicit life cycle GW and EU impacts at the county scale, with corn production in the U.S.'s Midwest region as a case study. This study compared the predictive accuracies and efficiencies of five distinct supervised machine learning (ML) algorithms, testing various sample sizes and feature selections. The results indicated that the gradient boosting regression tree model built with approximately 4000 records of monthly weather features yielded the highest predictive accuracy with cross-validation (CV) values of 0.8 for the life cycle GW impacts. The gradient boosting regression tree model built with nearly 6000 records of monthly weather features showed the highest predictive accuracy with CV values of 0.87 for the life cycle EU impacts based on all modeling scenarios. Moreover, predictive accuracy was improved at the cost of simulation time. The gradient boosting regression tree model required the longest training time. ML algorithms demonstrated to be one million times faster than the traditional process-based model with high predictive accuracy. This indicates that ML can serve as an alternative surrogate of process-based models to estimate life-cycle environmental impacts, capturing large geographic areas and timeframes.

**Keywords:** life cycle assessment; global warming; eutrophication; machine learning; spatial assessment; agriculture

## 1. Introduction

Meeting continuously increasing food and fuel demands while protecting environmental integrity is a grand challenge. Agriculture as the primary stage for food and fuel production is associated with a range of environmental pollution issues ranging from global warming to nutrient degradation.

Specifically, agriculture contributes to 8% of national greenhouse gases (GHGs) [1] and ranks as a leading contributor to nutrient pollution nationally and globally [2–4]. The estimated economic loss of environmental damage caused by nitrogen release alone already exceeds $210 billion/year in the United States (U.S.) [5]. In addition, the continuous increasing of food and fuel demands accompanied by population growth, energy and water shortages, and weather unpredictability, will further accelerate environmental pollution from agricultural expansion [6–10]. Ensuring environmentally friendly agricultural production is important to achieving multiple United Nations' Sustainable Development Goals, such as zero hunger, climate action, clean water and responsible production. Globally, food production ranks as a top contributor to water quality degradation in the form of eutrophication and hypoxia. To effectively mitigate agricultural pollution, it is urgent to accurately and rapidly assess the environmental impacts of agriculture.

The life cycle assessment (LCA) is the most prominent research method for quantifying environmental releases of agricultural supply chains [11–15]. Agricultural LCA is capable of quantifying environmental releases from both on-farm and supply chain activities, such as fertilizer production. The systematic scope of LCA is necessary in order to avoid shifting environmental burdens, such as between global warming and eutrophication impacts [16]. The recent LCA studies indicated that life cycle environmental releases of crop production present substantial spatial heterogeneity [17–20], as influenced by weather, soil and farming practices. For example, the reported life cycle nutrient releases of corn in the Midwest region varied by a factor of 200, spanning from 0.001 to 0.2 kg N-eq/kg of corn [13,17,20–29]. The spatially-explicit life cycle environmental impacts from agricultural production serve as a scientific basis for locating spatial hot-spots and designing spatially targeted remediation strategies.

Various approaches exist to estimating the life cycle environmental impacts of agricultural production [30,31]. However, it remains challenging to efficiently estimate, spatially and temporally, explicit life cycle environmental impacts at the county scale over a large region. The existing approaches belong to two general groups, based on their fundamental differences in the structure of life cycle inventory. The first group features the integration of spatially and temporally explicit on-farm releases from biogeochemical models and supply chain releases from process-based LCA. For crop systems, the existing crop LCAs utilized biogeochemical models such as the daily CENTURY (DAYCENT) model [1,29,32], environmental policy integrated climate model (EPIC) [13], and the denitrification-decomposition model (DNDC) [33,34] to estimate GHG and nutrient releases of corn [22,35], soybeans [35], alfalfa [35], hybrid poplar [35], switchgrass [22,35], and miscanthus [13,33]. These biogeochemical models are capable of capturing spatial explicit environmental releases, as influenced by location-specific weather, soil, topography and farming practices at finer spatial scales, such as farm or county scales [1,36,37]. Despite being powerful, these biogeochemical models are data-intensive and time-consuming, necessitate expert inputs and often require supercomputer clusters to compute the spatially explicit environmental impacts over a large region (i.e., the Midwest) [38,39]. The second group utilizes the input-output structure for the life cycle inventory. The multi-regional input-output LCA approaches, such as EORA and EXIOBASE [40], are utilized to track the environmental impacts of economic interdependencies among regions, often at a sector level. Although the input-output LCA approaches are capable of investigating the environmental impacts of trades over a large geographic area, the life cycle environmental inventory is often aggregated at a sector level and reside at coarse spatial scales, such as regional or county scales. EXIOBASE does not track the influences of temporally dynamic factors such as weather, soil and farming practices on life cycle environmental impacts. Overall, methods capable of rapidly and accurately estimating spatially and temporally an explicit life cycle environmental inventory of agricultural production at the county scale over a large region are lacking.

Machine learning (ML) approaches were applied to various disciplines, such as predicting traffic speed in urban places and identifying fertilizer recommendation classes. While these studies are valuable, they demonstrate applications of ML in LCA [41,42]. Machine learning (ML) approaches may present promising alternatives to traditional approaches for accurately and efficiently estimating spatially and temporally explicit LCA at the county scale over a large region across multiple years. Distinct from

traditionally process-based or input-output LCA approaches, ML models are capable of efficiently computing large datasets at the county scale across multiple years. Although ML techniques have promising potentials for transforming LCAs, the application of ML models to spatially and temporally LCA has been limited to date [43–48]. Previously, ML techniques were used to estimate either spatially generic life cycle release inventory of production processes [48,49], or life cycle toxicity impacts of chemicals [50]. A few recent studies applied neural networks to examine the life cycle GHG emissions associated with producing various crops based on literature sources or the spatially generic emission factors of the Intergovernmental Panel on Climate Change (IPCC) [6,7,51–56]. These valuable efforts greatly contributed to fusing LCA and ML. However, none of the existing studies assessed the possibility of using ML techniques for predicting spatially and temporally explicit life cycle environmental releases.

To fill in the aforementioned knowledge gaps, this study tested and compared five ML approaches for computing spatially and temporally explicit life cycle environmental impacts at the county scale, with corn production in the Midwest region as a case study. The research question this study addresses is which machine learning model will yield highest predictive accuracy and computational efficiency for estimating the spatially and temporally explicit life cycle environmental impacts of corn. The Midwest region contains more than 39 million hectares of land and produced more than 80 percent of U.S. corn in the year 2018 [57]. This study identified the most efficient the and accurate combination of ML approaches and datasets for predicting spatially explicit life cycle global warming (GW) and eutrophication (EU) impacts of corn production. To the authors' best knowledge, this is the first study applying and comparing ML approaches for predicting the spatially explicit life cycle environmental impacts of agricultural production.

## 2. Models and Data Sets

### *2.1. Overview of Models*

Five modeling approaches were tested and compared in this study, including linear regressison (LR), support vector machine regression (SVR) (Section 2.1.1), artificial neural network (ANN) (Section 2.1.2), two tree based model, gradient boosted regression tree (GBRT) (Section 2.1.3), and extreme gradient boosting (XGBoost) (Section 2.1.4) to determine the modeling approach which presented best predictive accuracy and the highest computational efficiency [58].

#### 2.1.1. Support Vector Machine Regression

Support vector machines (SVMs) are a class of learning based non-linear modeling algorithms with proven performance in a wide range of practical applications in classification and regression [59,60]. SVMs were initially developed for qualitative modeling problems. With the introduction of an $\epsilon$-insensitive loss function, SVMs have been extended to solve quantitative modeling problems. In this study, the SVM regression (SVR) technique was employed to model the county-level life cycle GW and EU impacts influenced by climate, soil characteristics, and farming practices features.

Suppose that a training set S containing $n$ data points is given as:

$$S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_i, y_i), ..., (\mathbf{x}_n, y_n)\} \quad (1)$$

where $\mathbf{x}_i \in R^d$, i.e., $d$-dimensional space, and $y_i \in R$. The goal of SVR is to find a function $f$ which can estimate all these data well. It is equivalent to the following constrained optimization problem:

$$\begin{aligned} \min \quad & \frac{1}{2}||w|| + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{s.t.} \quad & y_i - < w, x_i > -b \leq \epsilon + \xi_i \\ & < w, x_i > +b - y_i \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \end{aligned} \tag{2}$$

where $w$ is the normal vector of the hyperplane, $\xi_i$ and $\xi_i^*$ are the slack variables, and C is the coefficient of the regularization terms. In this case, it is set to equal to 1. $\mathbf{w}^T\mathbf{x}_i + b$ is the regression line which is the center of a tunnel with radius $\gamma = \frac{1}{||w||}$. Points on the surface so-called support vectors make up the tunnel. Involving $\xi_i, \xi_i^*$ allows left certain points outside tunnel as errors in order to avoid over-fitting. $\xi_i, \xi_i^*$ are set as 0.1.

The main purpose of SVR is to represent complicated relationships via non-linear mapping. The original input space is mapped into a high-dimensional feature space using kernel functions where it becomes linearly separable. Some common used kernel functions are linear, polynomial, sigmoid, and radial basis kernel functions (RBF) [61]. The RBF is the most often used kernel function in SVR.

$$k(x_i, x_j) = exp(-\gamma \parallel x_i - x_j \parallel^2) \tag{3}$$

The kernel function can be chosen according to data structure and the aim of project. Due to the small size of features, Gaussian kernel is used in this study.

Two parameters, *C* and $\gamma$ , must be appropriately set in SVR. The accuracy will be very high in the training stage and very low in the testing stage, when the value of C is set too large, while an extremely small value of may result in under-fitting, and excessively large value of may lead to over-fitting.

### 2.1.2. Artificial Neural Networks

*Artificial neural networks* (ANN) represent a type of computing that is based on the way that the human brain performs computations [62], which has been extensively used for pattern recognition and classification [63,64]. It consists of simple computational units called neurons. The numbers of neurons in the input and output layers are typically fixed by the type of application. The number of hidden layers and their neurons is typically determined by trial and error [65]. ANN can model complex non-linear relationships. The extra layers enable the composition of features from lower layers, giving the potential of modeling complex data with fewer units than a similarly performing shallow network [66].

The *multi-layer perceptron* (MLP) is the most frequently used ANN method, which consists of layered feedforward networks typically trained with static back propagation (see Figure 1).

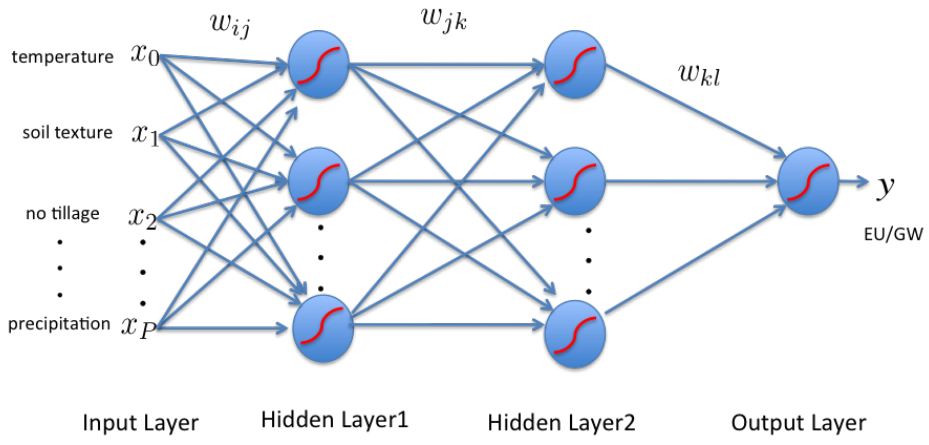


**Figure 1.** The architecture of MLP.

The neuron receives a set of inputs or signals (x), calculates a weighted average of them ($z$) using the summation function and weights ($w$), and then uses some activation function ($f$) to produce an output, where

$$z = \sum_{i=1}^{n} w_i x_i \tag{4}$$

The connections between the input layer and the hidden layer contain weights which are usually determined through training the system. The hidden layer sums the weighted inputs and uses the transfer function to create an output value. The transfer function is a relationship between the internal activation level of the neuron (called activation function) and the outputs. The rectified linear unit (ReLU) is the most commonly used activation function in deep learning models.

$$f(x) = x^+ = max(0, x) \tag{5}$$

where $x$ is the input to a neuron. The MLP was trained with back-propagation algorithm (BPA), the most frequently used in the ANN literature, and was adopted in this study.

The values of the weights are changed by using stochastic gradient descent (SGD) to minimize a suitable function used as the training stopping criterion. One of the functions most commonly used in supervised learning is the sum-of squared residuals given by Equation (6):

$$Loss = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{n} \tag{6}$$

where $y_i$ and $\hat{y}_i$ are the actual and predicted values, respectively.

The current weight change on a given layer is given by Equation (7).The weights will be updated by following the slope of the error surface downwards towards its minimum.

$$w_i^{new} \leftarrow w_i^{old} - \eta(\frac{\partial Loss}{\partial w^{old}}) \tag{7}$$

where $\eta$ (usually between $10^{-6}$ and 1.0) is the learning rate.

A high learning rate corresponds to rapid learning which may push the training towards a local minimum or cause oscillation. In turn, when applying small learning rates, the time to reach a global minimum will be considerably increase [67].

To achieve faster learning and avoid local minima, an additional term, called "momentum," is adopted to smooth out the weight changes, which helps to protect network learning from oscillation. To control the overfitting of a neural network, $L_2$ regularization and dropout are also used in this study.

#### 2.1.3. Gradient Boosting Regression Tree

*Gradient boosting regression tree* (GBRT) is an ensemble learning method that integrates the *regression tree* model with *gradient boosting* method. It builds a set of decision trees in a greedy manner at training time and makes a class prediction by majority voting. The algorithm is shown below.

1. Initialize

$$F_0(\mathbf{x}) = arg\min_{\gamma} \sum_{i=1}^{N} \mathbf{L}(y_i, \gamma) \tag{8}$$

   where L is the loss function, defined by Equation (10)

2. Compute

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}(x)} \tag{9}$$

   through $i = 1, ..., N$ and $m = 1, ..., M$. The $r_{im}$ are called pseudo-residuals.

3. Fit a regression tree to target $r_{im}$ giving terminal regions $R_{mj}$, for $j = 1, ..., J$ ($R_{mj}$ is a tree consisting of J leaf nodes).

$$\gamma_{jm} = arg \min_{\gamma} \sum_{x_i \in R_{mj}} L(y_i, f_{m-1}(x_i) + \gamma h_m(x_i)) \quad (10)$$

The new multiplier $\gamma_{jm}$ is updated based on the tree $f_{m-1}(x)$ and $\gamma_{jm-1}$.

4. The new tree is updated by

$$f_m(x) = f_{m-1}(x) + a \sum_{j=1}^{Jm} \gamma_{jm} I(x \in R_{mj}) \quad (11)$$

$I$ is the indicator function which equals to 1 if x is in the region $R_{mj}$. In this paper, we prevent over-fitting by controlling the learning rate and subsample. We set the learning rate to 0.01 and subsampling rate to 0.75.

5. Output $\hat{y}_i = F_M(x_i)$ by updating the following equation until $m$ reaches $M$.

$$f_M(x_i) = \sum_{m=1}^{M} f_m(x_i) = \sum_{m=1}^{M} \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{mj}) \quad (12)$$

The multiplier $\gamma_{jm}$ shrinks in small steps again to avoid over-fitting sufficiently. The maximum depth of the model is set to be 8 to constrain the function space. It controls the complexity of trees. In general, the GBRT model is voted by majority of M trees to decide the best step function to approximate the regression line. The total estimators is equal to 2000.

#### 2.1.4. Extreme Gradient Boosting

*Extreme gradient boosting* (XGBoost) is an improved algorithm based on the gradient boosting decision tree and can construct boosted trees efficiently and operate in parallel [68]. Compared with GBRT, XGBoost provides four additional features to achieve better performance. They are: (1) The weights of each new tree can be scaled down by given constant leaves. It reduces the influence of a single tree on the final score. (2) In column sampling, which works in a similar way to random forests, each tree is built using only a column-wise sample from the training dataset. (3) The introduction of the regularized loss function to avoid overfitting. (4) The loss function is approximated by Taylor expansion which speeds up the optimization procedure. We want to minimize the following regularized objective function in order to choose $f_k$.

$$\mathcal{L} = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (13)$$

$l$ is the loss function. In order to prevent too large a complexity of the model, the regularization term $\Omega$ is included as follows:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda ||w||^2 \quad (14)$$

where $\gamma$ is the complexity of each leaf. T is the number of leaves in the decision tree, and $\lambda$ is a parameter to scale the penalty which is set to be 0.85 in this study.

### *2.2. Datasets and Modeling Scenarios*

The predictor variables (Xs) included the features of temperature, precipitation, soil organic content, soil texture, application rates of nitrogen and phosphorus fertilizers, and farming practices. The outcomes variables (Ys) were defined as county-level life cycle GW and EU impacts. A total of 10 modeling scenarios reflecting various combinations of algorithms and features were tested and compared in this study (Table 1). Modeling scenarios S1A through S5A included total of 32 features, containing monthly average temperature and precipitation features. Modeling scenarios S1A through S5A included a total of 18 features, containing seasonally average temperature and precipitation features.

**Table 1.** A summary of modeling scenarios.

| Scenarios | Algorithms | Model Inputs (Features) |
|---|---|---|
| S1A | LM | monthly temperature (12 covariates), |
| S2A | SVR | monthly precipitation (12 covariates), |
| S3A | MLP | soil organic content, soil texture (3 covariates), |
| S4A | GBRT | Nitrogen fertilizer rate, phosphorus fertilizer rate, |
| S5A | XGBoost | percentage of conventional, no tillage (2 covariates) |
| S1B | LM | seasonal temperature (4 covariates), |
| S2B | SVR | seasonal precipitation (4 covariates), |
| S3B | MLP | soil organic content, soil texture (3 covariates), |
| S4B | GBRT | Nitrogen fertilizer rate, phosphorus fertilizer rate, |
| S5B | XGBoost | percentage of conventional, no tillage (2 covariates) |

The data on climate, soil characteristics, and farming practices were collected from various governmental databases for 12 U.S. Midwest states during the years 2000–2008 (Table 2). The monthly and seasonal average temperature (°C) and precipitation (mm) were obtained from the National Oceanic and Atmospheric Administration (NOAA)'s [69] National Climatic Data Center (NCDC) [70]. Soil data such as soil organic content and soil texture were extracted from the U.S. Department of Agriculture Soil Survey Geographic Database (SSURGO) [71]. The nitrogen and phosphorus fertilizer application rates and tillage practices were obtained from the U.S. Department of Agriculture's (USDA) [72] National Agricultural Survey Statistics and the Conservation Technology Information Center. The life cycle GW and EU impacts of corn production in Midwest states at the county-level during the years 2000-2008 are reported in our previous work (Lee et al. [17]). The total records are approximately 8000 over 9 years from 2000 through 2008. After removing records containing missing values, around 6000 complete records were left. To the authors' best knowledge, this is the largest database for spatially explicit life cycle impacts of corn production. The data sources for each feature are described in Table 2.

**Table 2.** Data sources for building the predictive model.

| Predictors (Xs) | Data Description | Data Sources |
|---|---|---|
| Temperature (°C) | Monthly mean temperature (Jan-Dec) | NOAA [69] |
| Precipitation (mm) | Monthly mean precipitation (Jan-Dec) | NOAA [69] |
| Soil organic content (%) | Percentage of soil organic content measured in soil depth up to 6 m | USDA SSURGO [71] |
| Soil type: Clay, Sand, Silt (%) | Percentage of soil types | USDA SSURGO [71] |
| Nitrogen and phosphorus fertilizers | fertilizer application rate (lbs/acre) | USDA NASS [72] |
| Farming practices (NT & CT) | Farming practices in fractions (No Tillage & Conventional Tillage) | USDA [72] |
| **Outcomes (Ys)** | | |
| Life cycle GW values | Life-cycle GW (kg $CO_2$-eq. kg $corn^{-1}$) | Lee et al. [17] |
| Life cycle EU values | Life-cycle EU (mg N-eq. kg $corn^{-1}$) | Lee et al. [17] |

### *2.3. Training, Validation, and Test Approaches*

The modeling procedures include four sequential steps: data preprocessing, hyper-parameter tuning, validation, and testing. The process is shown in Figure 2.

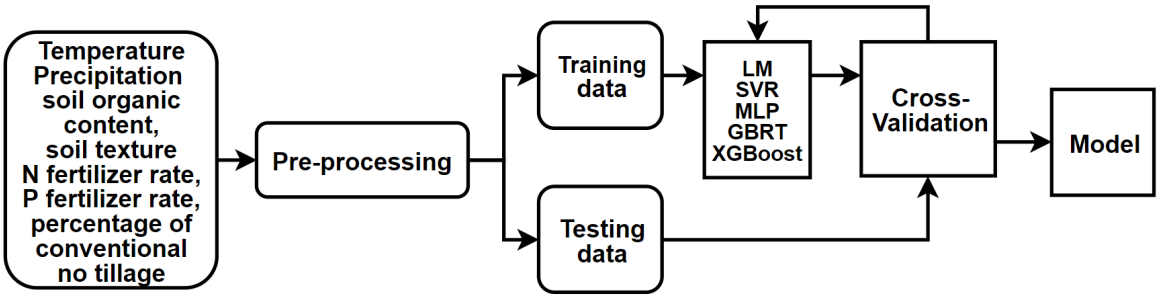


**Figure 2.** Key steps for constructing ML models.

To preprocess the raw data, we performed some trivial data cleaning, such as removing the missing values from the data using case-wise deletion. Categorical variables were one-hot encoded, while the numeric variables were normalized. The training step is equivalent to tuning the parameters for each model to seek the best fitted hyper-parameter combination. We call the parameters for the model hyper-parameters for distinguishing them from parameters of the data set. In order to find the best performing hyper-parameter combination, the grid search method was used. Two hyper-parameters, C and r, have to select the optimal value from two chooses of each kind. The grid, in this case, can be imagined as a 2 by 2 matrix. The names of hyper-parameters are taken to be the columns, and the two possible choices are taken to be the rows. The algorithm chooses one hyper-parameter from each kind at a time becoming a unique combination. Every model used in this study has different hyper-parameters. The different models should be assigned a distinct hyper-parameter grid. The grid search method would try every combination of preset parameters until it accessed all the combinations once. Preset parameters are given by experience. The algorithm would record every result from every hyper-parameter combination, and eventually return the best result and the related hyper-parameter combination. The best-performed parameter set would settle for the corresponding ML model.

After the first round of grid search, the rough selection provides a range of high performing hyper-parameter. This happens only if the hyper-parameters are continuous. For higher precision, the range is divided into several instances depending on the computer's power. In this study, we divided the hyper-parameters, such as the bag fraction for the GBRT model, into 10 instances.

In order to address the overfitting concern, we constructed models using a stricter 3-fold cross-validation methodology, in which we divided the three folds into training (two folds) and test (one fold) sets. The construction of the folds for cross-validation was created using the same stratification procedure. The process repeated three times until every fold was used as a validation fold once. The output from the 3-fold cross-validation would be the average of all folds of cross-validation results. Three metrics were used to evaluate the predictive accuracy of machine learning models, including cross-validation correlation, mean square error, and coefficient of determination:

(i) Cross-validation correlation coefficient (Corr, see Equation (15)),

$$corr = \frac{\sigma_{y_i}\sigma_{\hat{y}_i}}{\sigma_{y_i\hat{y}_i}} = \frac{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})} \quad (15)$$

was adopted to measure the strength of relationship between label $y_i$ and prediction $\hat{y}_i$.

(ii) Mean-square-error (MSE, see Equation (16))

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (16)$$

measures the average of the errors between $y_i$ and $\hat{y}_i$. This metric was involved to be the measure of training, testing, and cross-validation performance.

(iii) Coefficient of determination ($R^2$, see Equation (17))

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (17)$$

was used to evaluate how well the regression line represents the data. The computational efficiency was determined by the required time for training ML models and for using ML models for predicting spatially and temporally explicit life cycle environmental impacts, based on the computer configuration of Intel i7-4790k CPU, 8 GB memory at 2400 Mhz and Nvidia GTX 970 graphic card. Finally, we are able to calculate the scores under different metrics. The relative importance of input parameters such as weather, soil, and farming practices on life cycle GW and EU were ranked by sorting their corresponding coefficients. All the algorithms and metrics are adopted from Python package scikit-learn [73].

## 3. Results

### *3.1. Predictive Accuracy*

#### 3.1.1. Influences of Input Variables

The models built on monthly features generally yielded higher predictive accuracy than the models built on seasonal features for the same ML algorithms. For example, both cross-validation (CV) correlation and $R^2$ values of S1A (linear model based on monthly features) were larger than those of S1B (linear model based on seasonal features) for the life cycle GW. Similarly, both CV correlation and R-squared values of S1A were larger than those of S1B for the life cycle EU. The same trends of CV correlation and R-squared values were observed for the remaining modeling algorithms (Table 3), including XGBoost, GBRT, SVR, and MLP. Consistently, XGBoost and GBRT using monthly features achieved lower MSE values than their corresponding algorithms with seasonal features. The only exception was that LM, SVR, and MLP with monthly features presented slightly higher MSE values than their corresponding algorithms with seasonal features.

**Table 3.** Cross-validation (CV) correlation, $R^2$, and mean-square-error (MSE) values of all modeling scenarios.

| **Scenario A** | | **S1A LM** | **S2A SVR** | **S3A MLP** | **S4A GBRT** | **S5A XGBoost** |
|---|---|---|---|---|---|---|
| GW | CV corr | 0.45 | 0.68 | 0.64 | 0.80 | 0.78 |
| | MSE | 0.58 | 0.40 | 0.44 | 0.27 | 0.28 |
| | $R^2$ | 0.20 | 0.45 | 0.40 | 0.63 | 0.61 |
| EU | CV corr | 0.65 | 0.80 | 0.74 | 0.87 | 0.86 |
| | MSE | 20 | 14 | 18 | 10 | 9 |
| | $R^2$ | 0.42 | 0.60 | 0.50 | 0.75 | 0.73 |
| **Scenario B** | | **S1B LM** | **S2B SVR** | **S3B MLP** | **S4B GBRT** | **S5B XGBoost** |
| GW | CV corr | 0.35 | 0.63 | 0.64 | 0.78 | 0.76 |
| | MSE | 0.55 | 0.38 | 0.37 | 0.28 | 0.30 |
| | $R^2$ | 0.11 | 0.39 | 0.40 | 0.61 | 0.58 |
| EU | CV corr | 0.63 | 0.74 | 0.74 | 0.84 | 0.83 |
| | MSE | 22 | 17 | 18 | 10 | 11 |
| | $R^2$ | 0.39 | 0.53 | 0.49 | 0.71 | 0.62 |

### 3.1.2. Influences of Modeling Algorithms

Based on Table 3, GBRT presented the superior predictive accuracy. GBRT showed the highest CV correlation value (0.8) and R squared score (0.63), and the lowest MSE value (0.27) among all five models for the life cycle GW. Similarly, GBRT showed the highest CV correlation (0.87) and R squared score (0.75), and the lowest MSE value (10) among all five models for the life cycle EU. XGBoost was the second-best choice in terms of predictive accuracy. When the XGBoost model was applied for predicting life cycle GW, it produced the second-highest CV correlation value (0.78), the second-highest R squared score (0.61), and the lowest MSE value (0.28). Similarly, when the XGBoost model was used for predicting life cycle EU, it produced the second-highest CV correlation value (0.86), the second-highest R squared score (0.73), and the lowest MSE value (9). In contrast, the benchmark linear model exhibited the lowest predictive accuracy for both GW and EU. For example, the linear model had the lowest CV correlation value (0.45) and $R^2$ scores (0.2), and the highest MSE (0.58) among all five models for life cycle GW. Similarly, the linear model showed the lowest CV correlation (0.65) and R squared (0.63), and the lowest MSE (0.27) among all five models for the life cycle EU. Overall, XGBoost, GBRT, SVR, and MLP all showed higher predictive accuracy than the benchmark linear model, indicating that these models were capable of capturing non-linearity and complex interactions among features.

### 3.1.3. Influences of Sample Sizes

Figures 3 and 4 presented the influences of training data size on models' predictive capabilities. The CV correlation values for predicting life cycle GW exponentially increased for all model algorithms when the sample size increased from 100 to 4000. The CV correlation values for predicting life cycle GW reached the peak values at a sample size of round 4000. When the sample size continuously increased beyond 4000, the peak point, the CV correlation values slightly decreased. The CV correlation values for predicting life cycle EU dramatically increased for all model algorithms when the sample size increased from 100 to 5000. Additionally, the CV correlation values for GB and XGBoost exceeded 0.85 when 5000 input datasets were used to generate the models.

Figures 3 and 4 also confirmed that GBRT and XGBoost presented the highest predictive accuracy while the linear model always had the lowest predictive accuracy. For the same sample sizes across different algorithms, the GBRT model consistently showed the highest predictive accuracy among all models for both the life cycle GW and EU. The XGBoost model always ranked as the second-best fitted model, whose CV correlation values were slightly lower than the GBRT model under the same sample sizes.

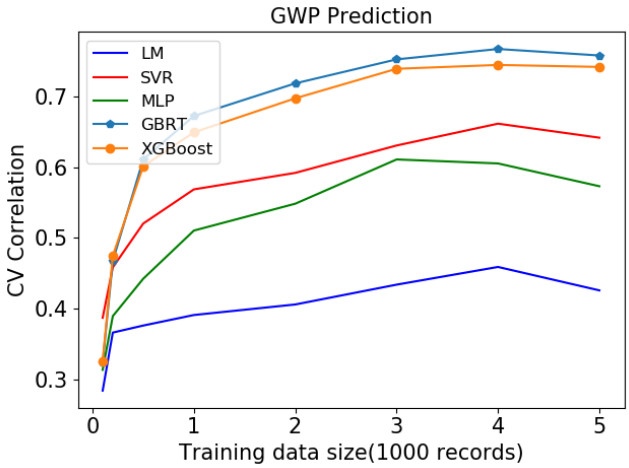


**Figure 3.** The influences of training sample sizes and algorithms on CV correlation values for life cycle global warming potential.

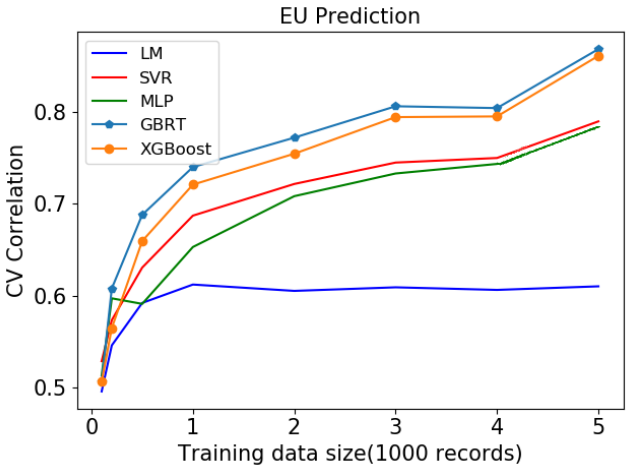


**Figure 4.** The influences of training sample sizes and algorithms on CV correlation values for life cycle eutrophication potential.

### *3.2. Predictive Efficiency*

All five models were a million times more efficient than the traditional process-based LCA models for estimating spatially explicit life cycle GW and EU. With the configuration of an Intel i7-4790k CPU, 8 GB memory at 2400 Mhz, and a Nvidia GTX 970 graphic card, the training time for five distinct algorithms ranged from 0.02 to 112 s for each simulation of hyper-parameter combination. While the linear model required the shortest duration of 0.02 s/simulation, GBRT needed the longest duration of 112 s/simulation for model training(Table 4). The predictive time for all models was less than 1 s. The total training and prediction times for all five models were less than 4 h for estimating spatially explicit life cycle GW and EU at the county scale in the entire Midwest region. In contrast, the traditional process-based LCA model approach relying on detailed bio-geochemistry models, such as EPIC, DNDC, and DAYCENT, may require months to estimate spatially explicit life cycle GW and EU for a large region such as the entire Midwest [17,39].

The comparison across five algorithms indicated trade-offs existed between computational efficiency and predictive accuracy for both the life cycle GW and EU. The linear model was the most efficient approach but presented the lowest predictive accuracy for both the life cycle GW and EU. In contrast, GBRT required the longest training time, but yielded the highest predictive accuracy for both impacts. In addition, although the time required for training GBRT for each hyper-parameter combination was three times longer than for training XGBoost, the CV correlation value for GBRT was only improved over 0.2 than for XGBoost.

**Table 4.** Time consumed for training and predicting each model in seconds.

| | | LM | SVR | MLP | GBRT | XGBoost |
|---|---|---|---|---|---|---|
| GW | Training | 0.02 | 4.11 | 65.4 | 112 | 33.8 |
| | Predicting | $7 \times 10^5$ | 0.13 | $1 \times 10^3$ | 0.09 | 0.06 |
| EU | Training | 0.03 | 4.72 | 33.1 | 128 | 41.6 |
| | Predicting | $6 \times 10^5$ | 0.04 | $5 \times 10^4$ | 0.02 | 0.09 |

### *3.3. Key Influential Factors*

The GBRT results showed that nitrogen and phosphorus application rates, soil organic content, and clay and silt types of soil were the top influencing factors for life cycle GW impacts (Figure 5). Nitrogen fertilizers contributed to life cycle GW impacts directly by causing on-farm nitrous oxide ($N_2O$) emissions from volatilization and denitrification processes in soil [74,75], and indirectly by emitting GHGs from supply chain activities such as manufacturing nitrogen fertilizers [76]. Phosphorus

fertilizers contributed to life cycle GW primarily by generating GHGs from supply chain activities such as manufacturing phosphorus fertilizers [76]. Moreover, soil characteristics also had a significant role in regulating GHG emissions. Recent studies based on field experiments [77,78] and mechanistic model (i.e., DNDC) [70] suggested that soil organic matter increased the availability of nitrogen and carbon to soil microorganisms, consequently enhancing $N_2O$ emissions. Furthermore, finer-textured soils, such as clay or silt soils, which tend to retain more water, often resulted in higher $N_2O$ emissions due to higher denitrification rates in soil [79]. Additionally, meteorological factors such as temperature also had a significant contribution to life cycle GW. For instance, March's temperature potentially influenced the shift of the $N_2O$:$N_2$ ratio [80], thereby increasing $N_2O$ emissions during the freeze-thaw cycle [81].

The present study also found that the top influencing factors of life cycle EU impacts were precipitation in February, soil organic content, the temperature in May, and precipitation in May and June (Figure 6). Precipitation can promote the penetration of nutrients through soil layers and wash away the nutrients from topsoil, thereby increasing nutrient leaching and runoff [82]. The contributions of precipitation to life cycle GW impacts were consistent with previous studies examining N fluxes in the watersheds of the contiguous U.S. [83,84] and Lake Michigan Basin [85], and field-based studies in the U.S. Midwest [86], where N losses to water bodies were higher during wet seasons/years compared to dry seasons/years. Furthermore, few studies [87,88] explained that soil organic carbon can lead to the accumulation of carbon and nutrients in the soil, subsequently resulting in more nutrients available for leaching and runoff, which corroborates the findings of this study. Amery and Vandecasteele also reported that lower phosphate binding capacity to soil organic portion may increase susceptibility to phosphorus leaching [89].

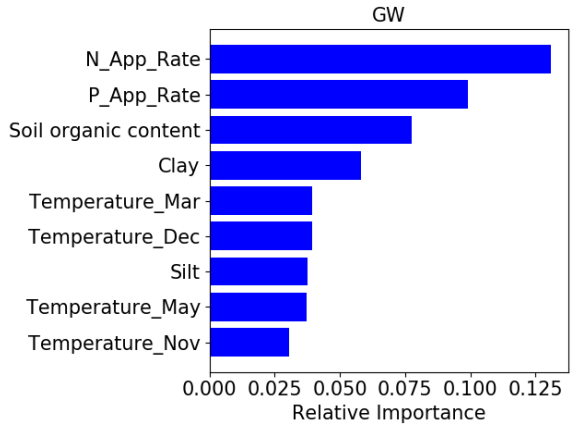


**Figure 5.** The top influential features for life cycle global warming potential.

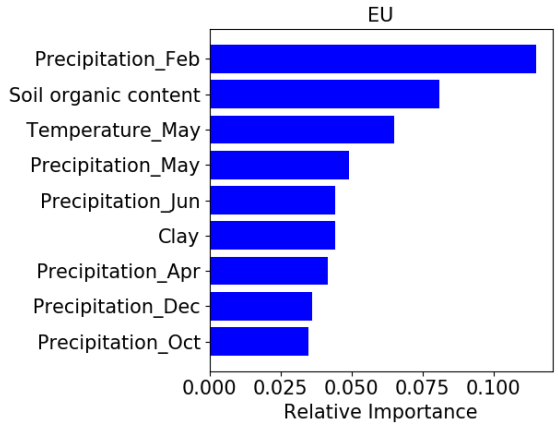


**Figure 6.** The top influential features for life cycle eutrophication potential.

## 4. Discussion

### *4.1. Syntheses of Machine Learning Modeling Comparison*

The comparison among all modeling scenarios indicated that the GBRT algorithm along with 4000 records of monthly features yielded the highest predictive accuracy for life cycle GW. The model based on the GBRT algorithm with 6000 records of monthly features showed the highest predictive accuracy for the life cycle EU. It is interesting to note that predictive accuracy was improved at the cost of simulation time. Although GBRT showed the highest predictive accuracy, GBRT required the longest training time. Compared with GBRT, the XGBoost model had slightly lower predictive accuracy but significantly lower computational time. Additionally, more training datasets (larger than 6000) will likely enable higher predictive accuracy for the life cycle EU. However, more training datasets (larger than 4000) cannot further improve the predictive accuracy for life cycle GW. Overall, when predictive accuracy is the key criterion, the GBRT model is the best choice. In contrast, when both predictive accuracy and model efficiency are top concerns, the XGBoost model is highly recommended. The majority of the existing machine learning applications in LCA used the ANN approach and did not compare performances between ANN and other ML approaches. However, our study found that BRT and XGBoost models showed higher predictive accuracy than the ANN approach for estimating corn's life cycle global warming and eutrophication impacts.

### *4.2. Application of Machine Learning Based Modeling Approaches for Supply Chain Management and Environmental Policies*

Compared with traditional process-based models, the ML models discussed in this study present unique merits, such as faster execution and reduced storage needs to estimate modeling outputs, and more flexible integration into other processes and simulation platforms. Such advantages permit ML models to rapidly complete a high number of simulation runs, thereby making ML models superior approaches for a range of computationally intensive tasks, such as optimization, prediction, and validation. For example, it is often infeasible for traditional process-based models to identify the optimal corn supply chains for corn-based biorefineries in the U.S. Midwest region at a fine spatial resolution over a large region, due to billions of simulation scenarios reflecting various combinations of weather, soil, farming practices, and supply scenarios. In contrast, the ML models compressed the model simulation time from months to seconds, consequently enabling the identification of the optimal corn supply chains for corn users such as biorefineries and animal sectors at a fine spatial resolution over a large geographic area.

The ML models in this study are valuable for policy assistance at the county, state, and regional scales. At the county and state scales, these models are capable of identifying the top polluting counties and states, and of suggesting spatially targeted remediation strategies. For example, different fertilizer application rates and tillage practices could be recommended for different counties to approximate an optimal trade-off level among life cycle GW and EU impacts. The results could also be aggregated into region scale assessments for supporting regional policies and planning. For example, to support regional biofuel and land use policies, our ML models could be used for corn supply chain optimization, corn-based biorefinery siting, and feedstock landscape optimization.

There are also several drawbacks associated with ML models. First, the computational burden for generating spatially explicit training datasets at fine spatial scales (such as a county) over a large geographic area (i.e., the Midwest) is substantial. The spatially explicit life cycle analyses at a fine spatial resolution over a large geographic area are lacking to date. The majority of LCAs that use spatially explicit biogeochemical models focused on a few farms or a small region due to data and computational constraints. This study obtained spatially explicit life cycle environmental impacts of corn at the county scale in the Midwest by integrating process-based LCA and the spatially explicit EPIC model with supercomputer clusters [17]. The computational burden for generating the spatially explicit training datasets likely will restrain the adoption of ML models, especially when the computing

infrastructure is not available for a large number of process-based modeling simulations. However, this study showed that increasing sample size beyond 4000 would not improve model accuracy for life cycle GW. This indicated that reducing the sample size to 4000 would be appropriate for reaching desirable predictive accuracy for life cycle GW, and is capable of reducing computational burden associated with generating training datasets and fitting ML models. We recommend future studies to identify the minimal sample size for spatially and temporally explicit LCAs of other crop and animal systems with machine learning approaches. Second, besides inheriting the uncertainty of process-based LCA and the EPIC model, the training algorithms also introduce additional uncertainty to our ML models. The uncertainty of process-based LCA and EPIC models was extensively discussed in previous studies [17,20]. This study focuses on the uncertainty introduced by training algorithms. By comparing four different algorithms and cross-validation procedures, this study found that the GBRT algorithm was the best statistical approach to approximate the traditional process-based model at a lower computational cost. We recommend future studies to identify the minimal sample size for spatially and temporally explicit LCAs of other crop and animal systems with machine learning approaches. Additionally, caution should be taken when applying the ML models generated from this study to other regions, whose weather, soil, or farming practice values may fall out of the ranges of our training datasets in the Midwest. In these cases, one should follow the procedures described in Figure 4 to construct and test ML models before employing them for other regions.

## 5. Conclusions and Recommendations for Future Work

While machine learning approaches were suggested as alternative approaches for predicting the life cycle environmental impacts of agricultural production, no studies yet identified the appropriate sample sizes and most suitable machine learning for predicting spatially and temporally explicit life cycle environmental impacts. This study comprehensively compared the influences of ML algorithms, training datasets, and sample sizes on the predicative accuracy and efficiency of ML models for estimating spatially explicit life cycle GW and EU impacts, with the Midwest corn as a case study. Among all LCA studies which used ML algorithms, this study included the largest training datasets, covering over 1000 counties in the Midwest over 9 years. This study found that GBRT and XGBoost models yielded the highest predictive accuracy for both the life cycle GW and EU. To further improve predictive accuracy, this study suggests expanding spatially and temporally explicit training datasets, particularly for spatially explicit life cycle EU. Moreover, the methods demonstrated in this study can be applied to other crop and animal systems. We recommend future studies to test these methods for predicting spatially explicit life cycle environmental impacts of soybean, beef, and dairy production, which are among the top contributors to the national GW and EU impacts. Additionally, the ML models open up new opportunities in solving computation-intensive challenges, such as optimization, prediction, and validation. Future interdisciplinary work, based on integrating the ML models in this study with optimization and economic models, will enable identifying the optimal supply chains, and supporting spatially and temporally explicit economic and policy analyses.

**Author Contributions:** Conceptualization, X.X.R.; data curation, X.X.R. and X.Z.; formal analysis, X.X.R., Z.G., and Y.P.; funding acquisition, X.X.R.; methodology, Z.G. and Y.P.; project administration, X.X.R.; resources, X.X.R.; supervision, X.X.R.; validation, X.X.R., Z.G. and Y.P.; visualization, Z.G.; writing—original draft, X.X.R.; writing—review and editing, X.X.R., Z.G., Y.P., E.K.L. and X.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This document is the result of the research project funded by Presidential Innovation Award at University at Albany, State University of New York.

**Conflicts of Interest:** The authors declare no conflict 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. Del Grosso, S.; Parton, W.; Mosier, A.; Walsh, M.; Ojima, D.; Thornton, P. DAYCENT national-scale simulations of nitrous oxide emissions from cropped soils in the United States. *J. Environ. Qual.* **2006**, *35*, 1451–1460. [CrossRef] [PubMed]
2. Bricker, S.; Longstaff, B.; Dennison, W.; Jones, A.; Boicourt, K.; Wicks, C.; Woerner, J. Effects of nutrient enrichment in the nation's estuaries: A decade of change. *Harmful Algea* **2008**, *8*, 21–32. [CrossRef]
3. Lassaletta, L.; Billen, G.; Grizzetti, B.; Garnier, J.; Leach, A.M.; Galloway, J.N. Food and feed trade as a driver in the global nitrogen cycle: 50-year trends. *Biogeochemistry* **2014**, *118*, 225–241. [CrossRef]
4. Council, N.R. *Clean Coastal Waters: Understanding and Reducing the Effects of Nutrient Pollution*; The National Academies Press: Washington, DC, USA, 2000.
5. Sobota, D.J.; Compton, J.E.; McCrackin, M.L.; Singh, S. Cost of reactive nitrogen release from human activities to the environment in the United States. *Environ. Res. Lett.* **2015**, *10*, 025006. [CrossRef]
6. Anyamba, A.; Small, J.; Britch, S.; Tucker, C.; Pak, E.; Reynolds, C.; Crutchfield, J.; Linthicum, K. Recent Weather Extremes and Impacts on Agricultural Production and Vector-Borne Disease Outbreak Patterns. *PLoS ONE* **2014**, *9*, e92538. [CrossRef]
7. Delcour, I.; Spanoghe, P.; Uyttendaele, M. Literature review: Impact of climate change on pesticide use. *Food Res. Int.* **2014**, *68*. [CrossRef]
8. Karmakar, R.; Das, I.; Dutta, D.; Rakshit, A. Potential Effects of Climate Change on Soil Properties: A Review. *Sci. Int.* **2016**, *4*, 51–73. [CrossRef]
9. Paerl, H.; Huisman, J. Climate Change: A Catalyst for Global Expansion of Harmful Cyanobacterial Blooms. *Environ. Microbiol. Rep.* **2009**, *1*, 27–37. [CrossRef]
10. Lee, E.K.; Zhang, W.J.; Zhang, X.; Alder, P.R.; Lin, S.; Feingold, B.J.; Khwaja, H.A.; Romeiko, X.X. Projecting life-cycle environmental impacts of corn production in the U.S. Midwest under future climate scenarios using a machine learning approach. *Sci. Total Environ.* **2020**, *714*, 136697. [CrossRef]
11. Smith, T.M.; Goodkind, A.L.; Kim, T.; Pelton, R.E.O.; Suh, K.; Schmitt, J. Subnational mobility and consumption-based environmental accounting of US corn in animal protein and ethanol supply chains. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 7891–7899. [CrossRef]
12. Xue, X.; Collinge, B.; Lanids, A.; Shrake, S.; Bilec, M. Regional life cycle assessment of soybean derived biodiesel for transportation fleet. *Energy Policy* **2012**, *48*, 295–303. [CrossRef]
13. Xue, X.; Landis, A.E. Eutrophication potential of food consumption patterns. *Environ. Sci. Technol.* **2010**, *44*, 6450–6456. [CrossRef] [PubMed]
14. ISO. Environmental Management and Life Cycle Assessment: Principles and Framework. Available online: https://www.iso.org/standard/37456.html (accessed on 20 May 2019).
15. ISO. Environmental Management and Life Cycle Assessment: Requirements and Guidelines. Available online: https://www.iso.org/standard/38498.html (accessed on 20 May 2019).
16. Notarnicola, B.; Sala, S.; Anton, A.; J.McLaren, S.; Saouter, E.; Sonesson, U. The role of life cycle assessment in supporting sustainable agri-food systems: A review of the challenges. *J. Clean. Prod.* **2017**, *140*, 399–409. [CrossRef]
17. Lee, E.K.; Zhang, X.; Adler, P.; Romeiko, X.X. Spatially and temporally explicit life cycle global warming, eutrophication, and acidification impacts from corn production in the U.S. Midwest. *J. Clean. Prod.* **2020**, *242*, 118465. [CrossRef]
18. Henderson, A.D.; Asselin-Balencon, A.; Heller, M.; Lessard, L.; Vionnet, S.; Jolliet, O. Spatial Variability and Uncertainty of Water Use Impacts from US Feed and Milk Production. *Environ. Sci. Technol.* **2017**, *51*, 2382–2391. [CrossRef]
19. Tabatabaie, S.M.H.; Bolte, J.P.; Murthy, G.S. A regional scale modeling framework combining biogeochemical model with life cycle and economic analysis for integrated assessment of cropping systems. *Sci. Total Environ.* **2018**, *625*, 428–439. [CrossRef]
20. Xue, X.; Pang, Y.; Landis, A. Evaluating agricultural management practices to improve the environmental footprint of corn-derived ethanol. *Renew. Energy* **2014**, *66*, 454–460. [CrossRef]
21. Adom, F.; Maes, A.; Workman, C.; Clayton-Nierderman, Z.; Thoma, G.; Shonnard, D. Regional carbon footprint analysis of dairy feeds for milk production in the USA. *Int. J. Life Cycle Assess.* **2012**, *17*, 520–534. [CrossRef]

22. Cronin, K.R.; Runge, T.M.; Zhang, X.; Izaurralde, R.C.; Reinemann, D.J.; Sinistore, J.C. Spatially Explicit Life Cycle Analysis of Cellulosic Ethanol Production Scenarios in Southwestern Michigan. *BioEnergy Res.* **2017**, *10*, 13–25. [CrossRef]
23. Grassini, P.; Cassman, K.G. High-yield maize with large net energy yield and small global warming intensity. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 1074–1079. [CrossRef]
24. Kim, S.; Dale, B.E. Life cycle assessment of fuel ethanol derived from corn grain via dry milling. *Bioresour. Technol.* **2008**, *99*, 5250–5260. [CrossRef] [PubMed]
25. Kim, S.; Dale, B.E. Cumulative Energy and Global Warming Impact from the Production of Biomass for Biobased Products. *J. Ind. Ecology* **2004**, *7*, 147–162. [CrossRef]
26. Kim, S.; Dale, B.E. Environmental aspects of ethanol derived from no-tilled corn grain: Nonrenewable energy consumption and greenhouse gas emissions. *Biomass Bioenergy* **2005**, *28*, 475–489. [CrossRef]
27. Landis, A.E.; Miller, S.A.; Theis, T.L. Life Cycle of the Corn-Soybean Agroecosystem for Biobased Production. *Environ. Sci. Technol.* **2007**, *41*, 1457–1464. [CrossRef]
28. Romeiko, X.X. A Comparative Life Cycle Assessment of Crop Systems Irrigated with the Groundwater and Reclaimed Water in Northern China. *Sustainability* **2019**, *11*, 2743. [CrossRef]
29. Xue, X.; Landis, A.E. Effect of agricultural practices on biofuels' environmental footprints. In Proceedings of the 2009 IEEE International Symposium on Sustainable Systems and Technology, ISSST'09, Phoenix, AZ, USA, 18–20 May 2009; pp. 1–4.
30. Gabrielle, B.; Gagnaire, N. Life-cycle assessment of straw use in bio-ethanol production: A case study based on biophysical modelling. *Biomass Bioenergy* **2008**, *32*, 431–441. [CrossRef]
31. Zaher, U.; Stockle, C.; Painter, K.; Higgins, S. Life cycle assessment of the potential carbon credit from no- and reduced-tillage winter wheat-based cropping systems in Eastern Washington State. *Agric. Syst.* **2013**, *122*, 73–78. [CrossRef]
32. Kim, S.; Dale, B.; Jenkins, R. Life cycle assessment of corn grain and corn stover in the United States. *Int. J. Life Cycle Assess.* **2009**, *14*, 160–174. [CrossRef]
33. Dufossé, K.; Gabrielle, B.; Drouet, J.L.; Bessou, C. Using Agroecosystem Modeling to Improve the Estimates of N2O Emissions in the Life-Cycle Assessment of Biofuels. *Waste Biomass Valorization* **2013**, *4*, 593–606. [CrossRef]
34. Gilhespy, S.L.; Anthony, S.; Cardenas, L.; Chadwick, D.; Prado, A.D.; Li, C.; Misselbrook, T.; Rees, R.M.; Salas, W.; Sanz-Cobena, A.; et al. First 20 years of DNDC (DeNitrification DeComposition): Model evolution. *Ecol. Model.* **2014**, *292*, 51–62. [CrossRef]
35. Adler, P.; Del Grosso, S.; Parton, W. Life-cycle assessment of net greenhouse-gas flux for bioenergy cropping systems. *Ecol. Appl.* **2007**, *17*, 675–691. [CrossRef] [PubMed]
36. J, P.; RJ, H.; WR., H. A metamodelling approach to estimate global n2o emissions from agricultural soils. *Glob. Ecol. Biogeogr.* **2014**, *23*, 912–924.
37. Zhang, X.; Izaurralde, R.; Manowitz, D.; West, T.; Post, M.; Thomson, A.; Bandaru, V.; Nichols, J.; Williams, J. An Integrated Modeling Framework to Evaluate the Productivity and Sustainability of Biofuel Crop Production. *Glob. Chang. Biol. Bioenergy* **2010**, *2*, 258–277. [CrossRef]
38. Necpálova, M.; Anex, R.P.; Fienen, M.N.; Grosso, S.J.; Castellano, M.J.; Sawyer, J.E.; Iqbal, J.; Pantoja, J.L.; Barkerd, D.W. Understanding the DayCent model: Calibration, sensitivity, and identifiability through inverse modeling. *Ecol. Model.* **2014**, *66*, 110–130. [CrossRef]
39. Nguyen, T.H.; Nong, D.; Paustian, K. Surrogate-based multi-objective optimization of management options for agricultural landscapes using artificial neural networks. *Ecol. Model.* **2019**, *400*, 1–13. [CrossRef]
40. Giljum, S.; Wieland, H.; Lutter, F.S.; Eisenmenger, N.; Schandl, H.; Owen, A. The impacts of data deviations between MRIO models on material footprints: A comparison of EXIOBASE, Eora, and ICIO. *J. Ind. Ecology* **2019**, *23*. [CrossRef]
41. Bratsas, C.; Koupidis, K.; Salanova Grau, J.M.; Giannakopoulos, K.; Kaloudis, A.; Aifadopoulou, G. A Comparison of Machine Learning Methods for the Prediction of Traffic Speed in Urban Places. *Sustainability* **2019**, *12*, 142. [CrossRef]
42. Garg, R.; Aggarwal, H.; Centobelli, P.; Cerchione, R. Extracting Knowledge from Big Data for Sustainability: A Comparison of Machine Learning Techniques. *Sustainability* **2019**, *11*, 6669. [CrossRef]

43. Dick, M.; Silva, M.A.d.; Dewes, H. Mitigation of environmental impacts of beef cattle production in southern Brazil e Evaluation using farm-based life cycle assessment. *J. Clean. Prod.* **2015**, *87*, 58–67. [CrossRef]
44. Marvuglia, A.; Kanevski, M.; Benetto, E. Machine learning for toxicity characterization of organic chemical emissions using USEtox database: Learning the structure of the input space. *Environ. Int.* **2015**, *83*, 72–85. [CrossRef]
45. Ramakrishnan, N.; Marwah, M.; Shah, A.; Patnaik, D.; Hossain, M.S.; Sundaravaradan, N.; Patel, C. Data Mining Solutions for Sustainability Problems. *IEEE Potentials* **2012**, *31*, 28–34. [CrossRef]
46. Slapnik, M.; Istenič, D.; Pintar, M.; Udovč, A. Extending life cycle assessment normalization factors and use of machine learning—A Slovenian case study. *Ecol. Indic.* **2015**, *50*, 161–172. [CrossRef]
47. Sousa, I.; Eisenhard, J.L.; Wallace, D. Approximate life-cycle assessment of product concepts using learning systems. *J. Ind. Ecology* **2001**, *4*, 61–81. [CrossRef]
48. Sundaravaradan, N.; Patnaik, D.; Ramakrishnan, N.; Marwah, M.; Shah, A. Discovering life cycle assessment trees from impact factor databases. In Proceedings of the Twenty-fifth AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 7–11 August 2011.
49. Hou, P.; Cai, J.; Qu, S.; Xu, M. Estimating Missing Unit Process Data in Life Cycle Assessment Using a Similarity-Based Approach. *Environ. Sci. Technol.* **2018**, *52*, 5259–5267. [CrossRef]
50. Song, R.; Keller, A.A.; Suh, S. Rapid Life-Cycle Impact Screening Using Artificial Neural Networks. *Environ. Sci. Technol.* **2017**, *51*, 10777–10785. [CrossRef]
51. Khoshnevisan, B.; Rafiee, S.; Omid, M.; Mousazadeh, H.; Rajaeifar, M. Application of artificial neural networks for prediction of output energy and GHG emissions in potato production in Iran. *Agric. Syst.* **2014**, *123*, 120–127. [CrossRef]
52. Khoshnevisan, B.; Rafiee, S.; Omid, M.; Yousefi, M.; Movahedi, M. Modeling of energy consumption and GHG (greenhouse gas) emissions in wheat production in Esfahan province of Iran using artificial neural networks. *Energy* **2013**, *52*, 333–338. [CrossRef]
53. Nabavi-Pelesaraei, A.; Rafiee, S.; Hosseinzadeh-Bandbafha, H.; Shamshirband, S. Modeling energy consumption and greenhouse gas emissions for kiwifruit production using artificial neural networks. *J. Clean. Prod.* **2016**, *133*, 924–931. [CrossRef]
54. Nabavi-Pelesaraei, A.; Rafiee, S.; Mohtasebi, S.; Hosseinzadeh-Bandbafha, H. Integration of artificial intelligence methods and life cycle assessment to predict energy output and environmental impacts of paddy production. *Sci. Total Environ.* **2018**, *631-632*, 1279–1294. [CrossRef]
55. Pahlavan, R.; Omid, M.; Akram, A. Energy input-output analysis and application of artificial neural networks for predicting greenhouse basil production. *Energy* **2012**, *37*, 171–176. [CrossRef]
56. Elhami, B.; Khanali, M.; Akram, A. Combined application of Artificial Neural Networks and life cycle assessment in lentil farming in Iran. *Inf. Process. Agric.* **2016**, *4*. [CrossRef]
57. USDA. Corn & Other Feedgrains. Available online: https://www.ers.usda.gov/topics/crops/corn-and-other-feedgrains/ (accessed on 20 October 2019).
58. Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer: Berlin/Heidelberg, Germany, 2006.
59. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
60. Schölkopf, B.; Smola, A.J.; Williamson, R.C.; Bartlett, P.L. New support vector algorithms. *Neural Comput.* **2000**, *12*, 1207–1245. [CrossRef] [PubMed]
61. Muller, K.R.; Mika, S.; Ratsch, G.; Tsuda, K.; Scholkopf, B. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.* **2001**, *12*, 181–201. [CrossRef]
62. Demuth, H.B.; Beale, M.H.; De Jess, O.; Hagan, M.T. *Neural Network Design*; Martin Hagan: Stillwater, OK, USA, 2014.
63. Palani, S.; Liong, S.Y.; Tkalich, P. An ANN application for water quality forecasting. *Mar. Pollut. Bull.* **2008**, *56*, 1586–1597. [CrossRef]
64. Kawabata, D.; Bandibas, J. Landslide susceptibility mapping using geological data, a DEM from ASTER images and an Artificial Neural Network (ANN). *Geomorphology* **2009**, *113*, 97–109. [CrossRef]
65. Gong, P. Geological mapping. *Photogramm. Eng. Rem. Sens.* **1996**, *62*, 513–523.

66. Bengio, Y.; others. Learning deep architectures for AI. *Found. Trends® Mach. Learn.* **2009**, *2*, 1–127. [CrossRef]
67. Topping, B.; Khan, A.; Bahreininejad, A. Parallel training of neural networks for finite element mesh decomposition. *Comput. Struct.* **1997**, *63*, 693–707. [CrossRef]
68. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD'16; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
69. NOAA. National Oceanic and Atmospheric Administration. Available online: https://www.noaa.gov (accessed on 20 May 2019).
70. NCDC. National Climatic Data Center. Available online: https://www.ncdc.noaa.gov (accessed on 20 May 2019).
71. USDA. Soil Survey Geographic (SSURGO) Database. Available online: https://sdmdataaccess.sc.egov.usda.gov (accessed on 20 May 2019).
72. USDA. National Agricultural Statistics Service: Quick Stats-Crop Yield Data. Available online: https://quickstats.nass.usda.gov (accessed on 20 May 2019).
73. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
74. Snyder, C.; Bruulsema, T.; Jensen, T.; Fixen, P. Review of Greenhouse Gas Emissions from Crop Production Systems and Fertilizer Management Effects. *Agric. Ecosyst. Environ.* **2009**, *133*, 247–266. [CrossRef]
75. Johnson, J.; Franzluebbers, A.; Weyers, S.; Reicosky, D. Agricultural opportunities to mitigate greenhouse gas emissions. *Environ. Pollut.* **2007**, *150*, 107–124. [CrossRef] [PubMed]
76. Ecoinvent. EcoInvent Data v2.2. Ecoinvent Reports No. 1-25. 2012. Available online: https://www.ecoinvent.org (accessed on 20 May 2019).
77. Stehfest, E.; Bouwman, A. N2O and NO emission from agricultural fields and soils under natural vegetation: Summarizing available measurement data and modeling of global annual emissions. *Nutr. Cycl. Agroecosyst.* **2006**, *74*, 207–228. [CrossRef]
78. Hoyle, F.; Barton, L.; Stefanova, K.; Murphy, D. Incorporating organic matter alters soil greenhouse gas emissions and increases grain yield in a semi-arid climate. *Agric. Ecosyst. Environ.* **2016**, *231*, 320–330. [CrossRef]
79. Adler, P.; Del Grosso, S.; Inman, D.; Jenkins, R.; Spatari, S.; Zhang, Y. Mitigation Opportunities for Life-Cycle Greenhouse Gas Emissions during Feedstock Production across Heterogeneous Landscapes. In *Managing Agricultural Greenhouse Gasses: Coordinated Agricultural Research through GRACEnet to Address Our Changing Climate*; Elsevier Inc.: New York, NY, USA, 2012; pp. 203–219. [CrossRef]
80. Butterbach-Bahl, K.; Dannenmann, M. Denitrification and associated soil N 2O emissions due to agricultural activities in a changing climate. *Curr. Opin. Environ. Sustain.* **2011**, *3*, 389–395. [CrossRef]
81. Congreves, K.; Wagner-Riddle, C.; Si, B.; Clough, T. Nitrous oxide emissions and biogeochemical responses to soil freezing-thawing and drying-wetting. *Soil Biol. Biochem.* **2018**, *117*, 5–15. [CrossRef]
82. Rabalais, N.; Diaz, R.; Levin, L.; Turner, R.; D, G.; J, Z. Dynamics and distribution of natural and human-caused hypoxia. *Biogeosciences* **2010**, *7*, 585–619. doi:10.5194/bg-7-585-2010. [CrossRef]
83. Sinha, E.; Michalak, A.; Balaji, V. Eutrophication will increase during the 21st century as a result of precipitation changes. *Science* **2017**, *357*, 405–408. [CrossRef]
84. Howarth, R.; Marino, R. Nitrogen as the Limiting Nutrient for Eutrophication in Coastal Marine Ecosystems: Evolving Views over Three Decades. *Limnol. Oceanogr.* **2006**, *51*, 364–376. [CrossRef]
85. Han, H.; Allan, J.D.; Scavia, D. Influence of Climate and Human Activities on the Relationship between Watershed Nitrogen Input and River Export. *Environ. Sci. Technol.* **2009**, *43*, 1916–1922. [CrossRef]
86. Gentry, L.; SILVER, T.; Below, F.; Royer, T.; Mcisaac, G. Nitrogen Mass Balance of a Tile-Drained Agricultural Watershed in East-Central Illinois. *J. Environ. Qual.* **2009**, *38*, 1841–1847. [CrossRef] [PubMed]
87. Wieder, W.; Cleveland, C.; Townsend, A. Throughfall exclusion and leaf litter addition drive higher rates of soil nitrous oxide emissions from a lowland wet tropical forest. *Glob. Chang. Biol.* **2011**, *17*, 3195–3207. [CrossRef]

88. Zhang, Y.; Wang, L.; Hu, Y.; Xi, X.; Tang, Y.; Chen, J.; Fu, X.; Sun, Y. Water Organic Pollution and Eutrophication Influence Soil Microbial Processes, Increasing Soil Respiration of Estuarine Wetlands: Site Study in Jiuduansha Wetland. *PLoS ONE* **2015**, *10*, e0126951. [CrossRef] [PubMed]
89. Amery, F.; Vandecasteele, B. Wat Weten We over Fosfor en Landbouw ? Beschikbaarheid van Fosfor in Bodem en Bemesting. Available online: https://www.vlaanderen.be/publicaties/wat-weten-we-over-fosfor-en-landbouw-deel-1-beschikbaarheid-van-fosfor-in-bodem-en-bemesting (accessed on 20 October 2019).