


APPLICATION PAPER

Data-driven decarbonization framework with machine learning

Ayush Jain¹ , Manikandan Padmanaban¹, Jagabondhu Hazra¹, Ranjini Guruprasad¹, Shantanu Godbole¹ and Heriansyah Syam²

¹IBM Research Labs India, Bengaluru, India

²Triputra Agro Persada, Jakarta, Indonesia

Corresponding author: Ayush Jain; Email: ayush.jain@ibm.com

A.J. and M.P. authors are contributed equally to this work.

Received: 12 January 2023; **Revised:** 12 April 2024; **Accepted:** 28 August 2024

Keywords: AI; carbon accounting; decarbonization; hotspot identification; palm oil supply chain

Abstract

Eight major supply chains contribute to more than 50% of the global greenhouse gas emissions (GHG). These supply chains range from raw materials to end-product manufacturing. Hence, it is critical to accurately estimate the carbon footprint of these supply chains, identify GHG hotspots, explain the factors that create the hotspots, and carry out what-if analysis to reduce the carbon footprint of supply chains. Towards this, we propose an enterprise decarbonization accelerator framework with a modular structure that automates carbon footprint estimation, identification of hotspots, explainability, and what-if analysis to recommend measures to reduce the carbon footprint of supply chains. To illustrate the working of the framework, we apply it to the cradle-to-gate extent of the palm oil supply chain of a leading palm oil producer. The framework identified that the farming stage is the hotspot in the considered supply chain. As the next level of analysis, the framework identified the hotspots in the farming stage and provided explainability on factors that created hotspots. We discuss the what-if scenarios and the recommendations generated by the framework to reduce the carbon footprint of the hotspots and the resulting impact on palm oil tree yield.

Impact Statement

Enterprises are under significant pressure from investors, consumers, and policymakers to act on climate change mitigation by disclosing their greenhouse gas emissions (GHG) emissions and committing to the reduction of emissions from their industrial activities. Over 20% of the world's largest companies have set long-term net-zero targets. Enterprises need technology to assess the carbon performance of their supply chain and receive informed recommendations for reducing their emission footprints. In this research work, we have proposed an AI framework to enable the stakeholders (private enterprises, suppliers, consumers, and service providers) with informed recommendations to decarbonize the supply chains. We have leveraged this framework to enable Palm producers to identify the factors responsible for higher GHG emissions and the actionable recommendation plans for reducing carbon footprints. For one palm field, it is observed that there is a scope to reduce the carbon emission by 20% by reducing the fertilizer application by 10% and manure application by 30%, with negligible impact on the yield.

1. Introduction

In the last few decades, supply chains have become more global, multi-echelon, inter-connected, and dynamic leading to benefits in terms of reducing costs, enhanced speed, diversifying operational sourcing,

and quality. However, these shifts bring with them the massive contribution of supply chains to greenhouse gas emissions (GHG) emissions. It is estimated that goods and services that are traded internationally contribute to about 22% of the global GHG emissions (GEI, 2016). Eight major supply chains are contributing to more than 50% of the global GHG emissions. These eight supply chains include food, construction, fashion, fast-moving consumer goods, electronics, auto, professional services, and other freight. Of these, food contributes to more than one-third followed by construction which contributes to 10% of the global GHG emissions (WNZ Challenge, 2021).

Enterprises in general particularly those who operate/are part of the above-mentioned supply chains are under significant pressure from investors, consumers, and policymakers to disclose their GHG emissions and commit to reducing emissions. Over 20 percent of the world's largest companies have set long-term net-zero targets (Black et al., 2021). To achieve net-zero targets, enterprises need technology to measure, track, and decarbonize (reduce their emissions) while building operational resiliency to the effects of climate change.

In this paper, we will discuss the novel framework/workflow called Enterprise Decarbonization Accelerator (EDA). EDA performs emission computation, hotspot identification with explainability, and what-if analysis to provide recommendations in an automated manner to accelerate the decarbonization journey of enterprises. To demonstrate the efficacy of the EDA in accelerating the decarbonization process, we will apply the EDA to the palm oil supply chain to measure carbon footprint, identify, and explain the factors causing the hotspots, and use what-if analysis to provide recommendations to mitigate carbon hotspots of the enterprise.

The rest of the paper is organized as follows. [Section 2](#) summarizes the related work. [Section 3](#) describes the EDA framework to estimate GHG emissions, identify hotspots, provide explainability, and perform what-if analysis to provide recommendations for the reduction/removal of hotspots. [Section 4](#) describes palm cultivation, the plantation data obtained from a leading producer of palm oil, and the application of EDA to the palm oil supply chain. [Section 5](#) presents the results obtained by applying the EDA to the palm oil supply chains. Finally, we summarize the novelty and relevance of the proposed EDA in the space of decarbonization in [Section 6](#) and present the concluding remarks in [Section 7](#).

2. Related work

Decarbonization of supply chains is an important area of study, which includes accounting of carbon footprints, identifying an inefficient process, understanding the factors that attribute to low performance, and recommending feasible intervenable actions for overall carbon reductions. This enables businesses to identify sustainability impacts across a range of attributes such as economic, environmental, social, and governance. It allows decision-makers to identify sustainability opportunities and prioritize reduction actions. In the contemporary literature, most of the decarbonization works revolve around carbon accounting and hotspot identification by using either qualitative (Kuhndt et al., 2002) or quantitative approaches (Rudnicka et al., 2017).

Hot spot analysis (HSA) (Kuhndt et al., 2002) is a qualitative approach that uses relative relevance numbers from existing studies to give a rough overview of relevant sustainability aspects. This approach has been used to identify hotspots in supply chains by comparing the relevant numbers. (Bienge et al., 2010) Utilized HSA to integrate social and environmental dimensions along the entire value chain and to identify relevant aspects for product-specific sustainability management. Norris et al. (2014) used the social hotspot database to study the social hotspots of numerous product categories, while Guo et al. (2020) conducted a global hotspot analysis concerning food loss and waste with its associated GHG emissions. HSA has also been used to analyze the environmental impacts of food supply chains, like Liedtke et al. (2010) performed HSA to identify resource-intensive hotspots in the life cycles of coffee and cream cheese. However, these qualitative approaches have limited applications as they do not provide any actual values for the impact factors but rather use relative relevance numbers to give a rough overview of relevant sustainability aspects.

Life cycle assessment (LCA) is a quantitative systems approach aimed at assessing the environmental impact of a product throughout its life cycle. Works like (Acquaye et al. (2011), Piringner et al. (2016), and Singh et al. (2015) utilized LCA to evaluate carbon dioxide equivalent (CO₂e) emissions and identify carbon hotspots in bio-diesel, maize silage, and beef supply chains, respectively. However, both HSA and LCA methodologies do not provide explainable insights into the identified hotspots. Also, they do not provide the stakeholders with recommendations that can help them in reducing their environmental impact. Attributional and consequential LCA (ALCA and CLCA) (Thomassen et al., 2008) address this gap to some extent. ALCA provides attribution of total emissions from the processes and material flows in a product life cycle, while CLCA provides information about the consequences of changes in the level of output of a product on the total emissions associated with the product. However, these approaches are limited to product life cycles and cannot be used for enterprise-level decarbonization.

As we will be evaluating the proposed framework for estimating the carbon footprint and hotspot identification for palm oil supply chains, we next discuss the related work in the space of palm oil supply chains. Several studies have focused on estimating the carbon footprint associated with the production of palm oil. Rodrigues et al. (2014) evaluated crude palm oil's GHG balance through an LCA approach, using average data from the Brazil region, while Subramaniam et al. (2020) assessed the water footprint of the palm oil supply chain. Few works have focused on specific parts of the cradle-to-gate extent of the supply chain, such as the LCA of oil palm seedlings (Muhamad et al., 2012) and transportation (Arshad et al., 2017). Most of the prior works on palm oil supply chain estimate carbon footprints through the LCA approach, using existing databases and process-specific emission data which are then used to identify carbon-intensive phases.

However, these works do not include explainability, (Lundberg and Lee, 2017) or what-if analysis and recommendations (Mothilal et al., 2020) that can be helpful for emission reduction. In this work, we seek to address this gap by proposing enterprise decarbonization accelerator (EDA), a novel framework that performs emission computation, hotspot identification with explainability insights, and what-if analysis to recommend intervenable measures for reducing carbon footprint and help enterprises accelerate their decarbonization journey.

3. Decarbonization accelerator framework

We have designed and developed an enterprise scale decarbonization accelerator (EDA) framework (as shown in Figure 1) that would be able to perform the processes associated with emission computation, hotspot identification with explainability, and what-if analysis to provide recommendations in an automated manner to accelerate the decarbonization journey of enterprises. The proposed EDA framework consists of an AI workflow with four modules that is carbon accounting, carbon hotspot

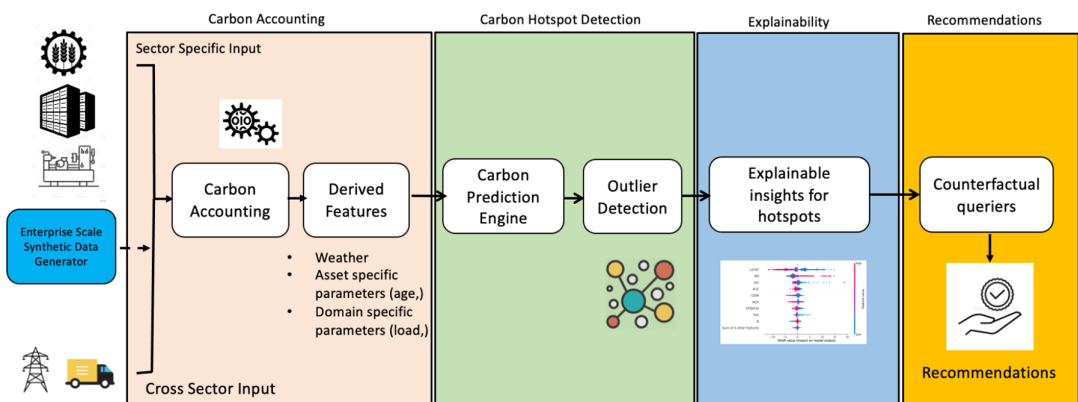


Figure 1. Enterprise decarbonization accelerator framework.

identification, explainability, counterfactual queries, and recommendations engine. Next, we describe the four modules.

3.1. Carbon accounting

GHG protocol, initiated by the World Resources Institute and the World Business Council for Sustainable Development, serves as a prominent framework for corporate accounting and reporting emissions. The GHG Protocol Corporate Accounting and Reporting Standard (Protocol and Initiative, 2004) categorizes greenhouse gas emissions into three scopes: Scope 1, encompassing direct emissions from owned or controlled sources; Scope 2, involving indirect emissions from purchased energy; and Scope 3, addressing indirect emissions along the entire value chain. We have implemented Application Programming Interface (APIs) for the scalable computation of Scopes 1, 2, and 3 emissions as per the GHG protocol.

3.1.1. Scope1: stationary combustion

Combustion of fuels in stationary (non-transport) combustion sources results in the following GHG emissions: CO₂, methane (CH₄), and nitrous oxide (N₂O). Sources of emissions from stationary combustion include boilers, heaters, furnaces, kilns, ovens, flares, thermal oxidizers, dryers, and any other equipment or machinery that combusts carbon-bearing fuels or waste stream materials. We compute the GHG emissions for the above gases for a variety of fuels using our API.

3.1.2. Scope1: mobile combustion

Mobile combustion means emissions from the transportation of materials, products, waste, and employees resulting from the combustion of fuels in company-owned or controlled mobile combustion sources (e.g., cars, trucks, buses, trains, airplanes, ships, etc.). The greenhouse gases CO₂, CH₄, and N₂O are emitted during the combustion of fuels in mobile sources. As per the GHG protocol, we have adopted a fuel-based approach for accounting carbon footprint from mobile combustion. We consider location, vehicle type, fuel type, and amount as input to mobile combustion API.

3.1.3. Scope1: fugitive emission

Fugitive emissions are leaks and other irregular releases of gases or vapors from a pressurized containment—such as appliances, storage tanks, pipelines, wells, or other pieces of equipment—mostly from industrial activities. We have leveraged a sales-based approach for computing the carbon footprint for fugitive emissions in our API.

While stationary, mobile, and fugitive emissions are common Scope 1 activities across enterprises, Scope 1 also includes industry-specific activities such as applying fertilizer to agricultural fields or production of Ozone-depleting gases. In such cases, our framework has the flexibility to include industry-specific Scope 1 carbon accounting models.

3.1.4. Scope2: emission from electricity consumption

Scope 2 API accounts for GHG emissions from the generation of purchased electricity consumed by a company. Purchased electricity is defined as electricity that is purchased or otherwise brought into the organizational boundary of the company. Scope 2 emissions physically occur at the facility where electricity is generated and attributed to the users based on their consumption. There are two approaches for Scope 2 that is location-based approach and the market-based approach. For simplicity, in this paper, we used the location-based approach.

3.1.5. Scope3

Scope 3 emissions are a consequence of the activities of the company but occur from sources not owned or controlled by the company. Some examples of Scope 3 activities are extraction and production of

purchased materials; transportation of purchased fuels; and use of products and services. In this paper, we used Scope 3 API to compute the carbon footprint for logistics using the weight-distance method.

This module also extracts carbon performance-related features such as weather conditions (e.g., temperature, humidity), asset-specific parameters (e.g., age, size–capacity), and operational parameters (e.g., load, fuel–electricity consumption) associated with the carbon performance of the asset/operation. While weather data is pulled automatically based on location information, users need to define the asset-specific parameters as well as operational parameters as domain-specific knowledge is required here.

3.2. Carbon hotspot detection

The carbon hotspot detection module ingests the carbon footprints of assets along with other relevant associated derived features and selects the prediction and outlier detection algorithm from the library of options, such as linear regressor, decision tree, random forest, Multilayer Perceptron (MLP) regressor, gradient boosting, XGBoost, PyOD, isolation forest, etc., to identify the list of low carbon performing assets/operations. It mainly consists of two sub-modules, namely, (i) emission prediction model and (ii) outlier (note that we use the terms, outlier, anomalous, and hotspot interchangeably) detection model as follows:

3.2.1. Emission prediction model

We learn the functional relationship between the carbon footprint of an asset and other independent associated derived factors by selecting the prediction model configured by the user from the library. We provide well-known ML models as part of our framework (such as Linear regression, Decision tree, Random Forest, Gradient boosting, and MLP Regressor). These models are leveraged from the sklearn python library (Pedregosa et al., 2011). Our framework performs emission modeling using the different models configured by the user and selects the optimal model using hyperparameter tuning.

3.2.2. Outlier detection model

In this module, we support out-of-the-box outlier detection modules (e.g., PyOD, Isolation Forest, etc.) as well as advanced methods such as multi-dimensional subset scan (MDSS). We used an extension of the MDSS (Neill et al., 2013; Zhang and Neill, 2016) algorithm for subset scanning of the data. Figure 2

Algorithm 1 Pseudocode for Multi-dimensional Subset Scan

```

Initialize best_score, i, cur_subgroup;
cur_data_subset = Data|cur_subgroup
repeat
  1. Randomly order the given m features to scan from 1 to M
  for j = 1 to M do
    1. cur_data_subset = Data|cur_subgroup-j
      (relax the subgroup definition to include all values of feature j)
    2. cur_subgroup = MDSS(cur_dataset)
      (Use MDSS on cur_dataset to identify the exact highest scoring
subset of values of feature j, given cur_subgroup-j)
    3. cur_data_subset = Data|cur_subgroup
    4. best_score = scorebias(cur_dataset)
  end for
  2. Check end condition, else loop through features in random order again,
i = i + 1
until best_score has not changed between i and i – 1

```

Figure 2. MDSS pseudocode.

shows the algorithm provided in Zhang and Neill (2016). This methodology is able to identify the most anomalous subgroup of feature space in linear time, amongst the exponentially many possible ones, enabling tractable subgroup analysis. The general form of the method is

$$S^* = \text{MDSS}(\mathbb{D}, \mathbb{E}, \text{score}_{\text{bias}}) S^* = \text{MDSS}(\mathbb{D}, \mathbb{E}, \text{score}_{\text{bias}}) \quad (3.1)$$

where S^* is the most anomalous subgroup, \mathbb{D} is a dataset with outcomes Y and features \mathbb{X} , \mathbb{E} are a set of expectations for Y , and $\text{score}_{\text{bias}}$ is an expectation-based scoring statistic that measures the amount of anomalousness between subgroup observations and their expectations. The goal of MDSS is to identify a subset of the data $d(S) \subseteq \mathbb{D}$ corresponding to subgroup S that maximizes $\text{score}_{\text{bias}}$. For $\text{score}_{\text{bias}}$, we use log-likelihood ratio defined as $F(S) = \log(\text{Pr}(D|H_1(S))/\text{Pr}(D|H_0))$. The alternate hypothesis $H_1(S)$ assumes that datapoints $x_i \in d(S)$ are drawn with mean $q\mu_i$ and datapoints $x_i \notin d(S)$ are drawn from mean μ_i , for constant multiplicative factor $q > 1$ known as relative risk. The null hypothesis H_0 assumes that all datapoints, including $x_i \in d(S)$ are drawn with mean μ_i . This definition satisfies the additive linear-time subset scanning property, which is required for MDSS to be tractable (Speakman et al., 2016). Equation (3.2) gives the general $\text{score}_{\text{bias}}$ used by MDSS.

$$\text{score}_{\text{bias}}(S) = \max_{q>1} \sum_{x_i \in d(S)} (\log \text{Pr}(x_i|q\mu_i) - \log \text{Pr}(x_i|\mu_i)) \quad (3.2)$$

3.3. Explainability

The interpretation of machine learning models is crucial for understanding feature attribution, especially as many models are inherently not transparent. In this module, we employ the widely recognized model-agnostic methodology known as SHapely Additive exPlanations (SHAP) (Lundberg and Lee, 2017) to gain insights into both local and global interpretability for carbon hotspots. SHAP is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions. The SHAP values calculated in our framework represent the marginal contribution of each associated feature to the carbon footprint of an asset. To compute these SHAP values, we leverage the SHAP library available at <https://pypi.org/project/shap/>. This approach ensures a robust and mathematically sound representation of feature importance. Moreover, our module is designed with a pluggable framework, allowing seamless integration of other explainable models into the workflow.

3.4. Counterfactual queries and recommendation engine

In this module, we provide a framework to leverage the state-of-the-art as well as custom counterfactual queries and recommendation engines. This framework has an inbuilt recommendation module that leverages the diverse counterfactual explanations (DiCE) methodology from Mothilal et al. (2020). DiCE generates counterfactual explanations for any ML model through perturbations within a feasible range that change the output of a machine learning model. It also supports simple constraints on features to ensure the feasibility of the generated counterfactual examples. This framework takes input from the user on the set of controllable features with their feasible range and the expected target emissions for intervenable actions and generates the set of best recommendations.

4. Case study: palm cultivation

Sustainable sourcing of palm oil has gained significant interest over the past years. It is thus important for palm producers to identify factors responsible for higher emissions, while also developing explainable intervention plans to lower their carbon footprint. We use our EDA framework to estimate GHG emissions of palm plantations, perform explainable hotspot identification, and develop recommendation plans for reducing carbon emissions.

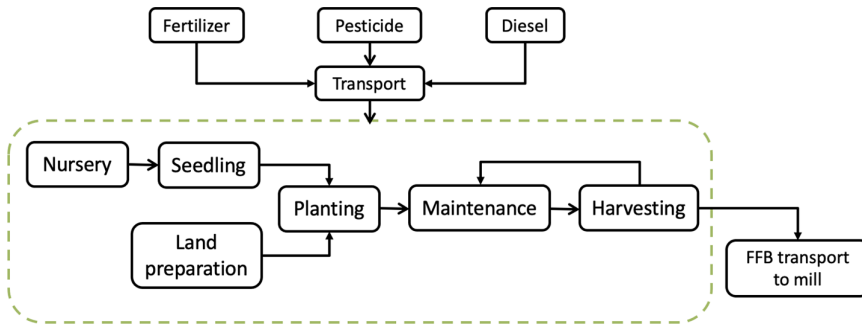


Figure 3. Palm cultivation: overview.

4.1. Overview

Figure 3 gives an overview of palm cultivation. The first step is to produce seedlings from pre-germinated seeds in nurseries, along with land preparation. Then the seedlings are planted manually and fertilizers containing elements such as nitrogen (N), potassium (K), and phosphorus (P) are applied. Fertilizers used are manufactured in factories from where they are taken to the port nearest to the plantations via shipping tankers. Then they are delivered by heavy goods vehicles (HGV) such as trucks to the plantations. Along with fertilizers, diesel is also transported to the plantations, where it is used in irrigation, transportation of fertilizers and workers to the farms, land preparation, and maintenance. It takes about 3–4 years for oil palms to produce fruits suitable for harvest. Palm trees continue to produce fruit for around 30 years and they are harvested periodically. Palm is harvested manually wherein fronds are cut off to dislodge fresh fruit bunches (FFB), which fall to the ground and are then collected. The fruit bunches are transported to oil mills by light good vehicles, where they are subjected to industrial processing to obtain crude palm oil (CPO) and crude palm kernel oil, as well as by-products such as empty fruit bunches (EFB), fibers, shells and palm oil mill effluent (POME). These by-products are returned to the field as manure. Electricity is used in the oil extraction phase to power machinery such as threshers and motors of conveyors.

GHG emissions from the palm supply chain can be divided into the following stages—manufacturing, agriculture, and transportation and electricity usage. Manufacturing includes all the emissions resulting from the production of fertilizers and electricity and the extraction of crude palm oil. Agriculture includes all the emissions resulting from activities related to the planting and harvesting of palm produce. Transportation includes emissions resulting from transporting fertilizers and diesel to palm plantations, and FFB produce from plantations to mills. Lastly, emissions due to electricity usage include all the indirect emissions resulting from the consumption of electricity in the supply chain.

4.2. Palm plantation data

Palm plantations are divided into smaller units called blocks. We use our framework to analyze the performance of palm plantations at the block level, using data from 25 palm blocks for 14 years, from the year of the plantation, up to the 14th year of cultivation of each block. We consider important factors which impact palm cultivation and capture the carbon emission performance of blocks, such as nitrogen content in the fertilizer application, carbon and nitrogen content in manure application (EFBh and pruned fronds), age of the farm block, initial soil organic carbon content at the depth of 0–15 cm and 15–45 cm, annual yield and the weather parameters (annual precipitation and temperature statistics). Data is obtained from a leading producer of palm oil. For the sake of anonymity, we will not disclose the name of the palm oil producer. Along with the farm plantation data, we also use relevant data from the cradle-to-gate extent of the palm oil supply chain for other stages - manufacturing, transportation, and electricity.

Figure 4 shows the heatmap of the annual yield of the 25 blocks at different ages, where the yield of each block has been normalized using min-max scaling. Since palm plantations produce their first harvest in the third–fourth year after plantation, the yield is near zero for the first 3 years across all the blocks. We

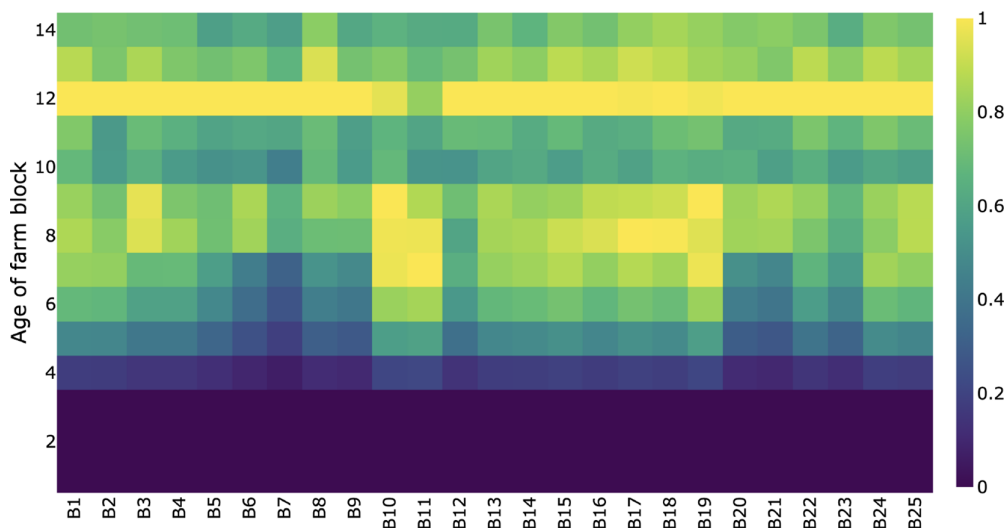


Figure 4. Heatmap showing normalized annual yield from the farm blocks at different ages.

see that the farm blocks have more yield as their age increases, with maximum yield observed around the 12th year of plantation across most of the blocks.

5. Experimentation and results

In this section, we discuss the results that are obtained by following the methodology outlined in Section 3 for palm plantation blocks and draw useful insights.

5.1. Carbon footprint of palm blocks

We have used the physics-based denitrification decomposition (DNDC) model (Gilhespy et al., 2014) to estimate the carbon footprint of palm plantations. The model bridges the chemical reactions in soil with the ecological drivers (climate, soil, vegetation, and farming practices) and environmental factors (temperature, moisture, pH, redox potential (Eh), and simulates carbon and nitrogen dynamics. The DNDC model predicts crop growth, soil temperature and moisture, soil carbon sequestration, and emission of CO₂, CH₄, and N₂O along with other trace gases.

Figure 5 shows the inputs, outputs, and components of the model. The inputs to the model are the soil data (type, clay fraction, amount of initial soil organic carbon), weather data (temperature, precipitation, wind speed, solar irradiation), crop data, farming practices (fertilizer, manure, and irrigation schedules, tillage information), planting and harvesting schedules. The two-component DNDC model has six submodules, namely, soil climate, crop growth, denitrification and decomposition, nitrification, and fermentation.

The soil climate sub-model computes environmental factors within the soil, such as soil temperature, moisture profiles (0–50 cm), pH, and redox potential. These calculations are based on soil texture properties, air temperature, precipitation, and plant water uptake. In parallel, the plant growth sub-model predicts daily water and nitrogen uptake by plants, as well as daily growth, root respiration, and biomass partitioning into grain, stems, and roots. The decomposition model integrates inputs from the soil climate and plant growth models to determine the quantities of substrates, including ammonium (NH₄⁺), nitrates (NO₃⁻), and dissolved organic carbon.

Depending on the soil moisture profile and redox potential levels, one of the nitrification, denitrification, or fermentation sub-models is activated. Based on the level of soil moisture profile and redox potential, any one of the sub-models of nitrification, denitrification, or fermentation will be invoked. Under aerobic conditions, the nitrification sub-model calculates the nitrification process, leading to the

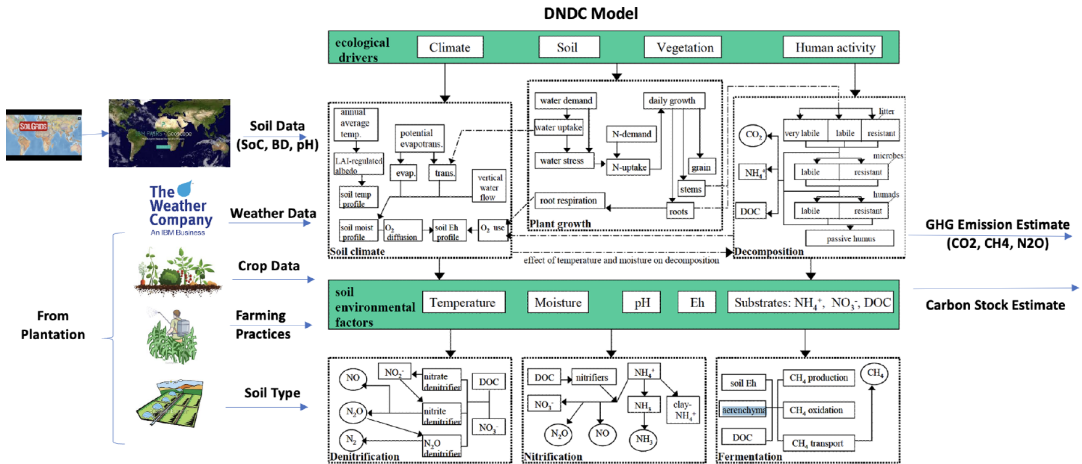


Figure 5. Denitrification-decomposition (DNDC) model workflow (Gilhespy et al., 2014).

production of N₂O, nitric oxide, and (NH₃). Conversely, during anaerobic conditions, the denitrification and/or fermentation sub-models estimate the emission of N₂O and CH₄, respectively.

The residual crop material integrated into the soil will be categorized into three distinct soil litter pools—namely, highly labile, labile, and resistant litter pools based on their carbon-to-nitrogen (C/N) ratio. This incorporation of litter serves as vital input for the storage of soil organic matter (SOM), thereby connecting the plant and soil components within a biogeochemical system. Within the DNDC framework, SOM is distributed across four primary pools: plant residue (i.e., litter), microbial biomass, humads (i.e., active humus), and passive humus. Each pool comprises two or three sub-pools characterized by varying specific decomposition rates. The daily decomposition rate for each sub-pool is governed by factors such as pool size, specific decomposition rate, soil clay content, nitrogen availability, soil temperature, and soil moisture. As the soil organic carbon (SOC) within a pool undergoes decomposition, a portion of the carbon is released as CO₂, while the remainder is allocated to other SOC pools. It is important to note that our study focuses solely on greenhouse gas (GHG) emissions from farming activities, and we do not consider changes in soil organic matter.

This DNDC model has been widely accepted and analyzed by the researchers in last two decades, and several validations with field measurements have been studied for 60+ crops in various countries (Gilhespy et al., 2014). For example, a field validation experiment of DNDC model for CH₄ and NO_x emissions from paddy fields is studied in India and it is reported that the DNDC model satisfactorily simulated the seasonal variations in GHG emissions for different land management (Babu et al., 2006).

Our framework computes carbon footprint at each stage from cradle to gate, that is raw material extraction to CPO production leveraging the APIs mentioned in Section 3.1 and the DNDC model. Emissions related to activities within the organizational boundary were computed using Scope1, Scope2 APIs, and DNDC model. Upstream emissions (i.e., raw material extraction and manufacturing of fertilizer) and emissions related to the transportation of upstream raw material and fertilizer were computed using the Scope3 logistic API. At the farming stage, the DNDC tool has been used to precisely compute the carbon footprint due to farming activities. The framework uses “CO₂e” as the standard unit for measuring carbon footprint. For any quantity and type of greenhouse gas, CO₂e signifies the amount of CO₂ which would have the equivalent global warming impact. A quantity of GHG can be expressed as CO₂e by multiplying the amount of the GHG by its global warming potential (GWP), where GWP is the heat absorbed by any greenhouse gas in the atmosphere, as a multiple of the heat that would be absorbed by the same mass of CO₂. For example, the GWP value for N₂O is provided as 265 in the IPCC Fifth Assessment Report, 2014 (AR5) (Myhre et al., 2013; Myhre et al., 2014). Therefore, if 1 kg of N₂O is emitted, this can be expressed as 265 kg of CO₂e (1 kg N₂O * 265 = 265 kg CO₂e).

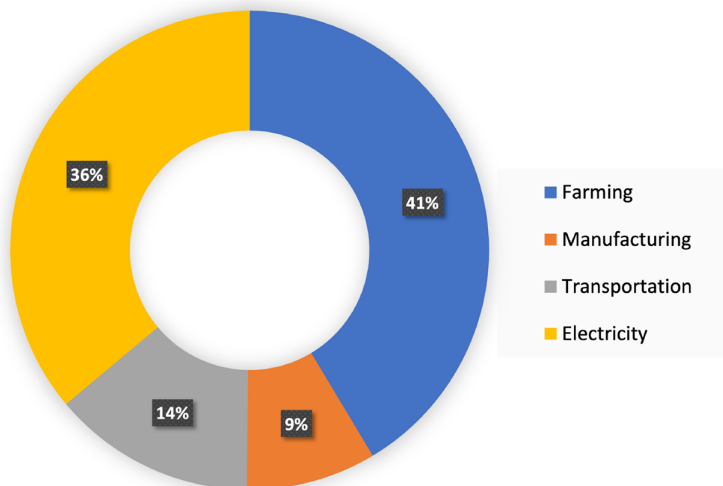


Figure 6. Stage-wise distribution of carbon emission.

The emissions were divided into four stages—farming, manufacturing, transportation, and electricity as mentioned in Section 4.1. Figure 6 shows the stage-wise distribution of carbon emissions in the cradle-to-gate palm oil supply chain, averaged across the 14 years of plantation data. The plot shows that the farming stage is the major source of emissions in the supply chain. Therefore, in this paper, we will focus on the farming stage, and use our framework for block-level analysis of palm plantation.

The DNDC model gives the emission of GHG gases such as CO₂, CH₄, and N₂O from the farming stage at daily intervals. For our analysis, we have aggregated the GHG emissions to annual temporal resolution. Since all the palm plantation blocks are located in the upland mineral soil, the methane emission from the farming stage is negligible. We have considered only the soil CO₂ emission from the decomposition process and N₂O emission from the nitrification and denitrification process to compute the total carbon footprints (CO₂e).

Figure 7 shows the heatmap of annual CO₂e emission of the 25 palm plantation blocks at different ages, where the emission numbers are normalized across all the blocks using min–max scaling. All 25 blocks receive an equal amount of fertilizers for the first 4 years and for subsequent years, the amount of fertilizers and manures are determined based on the age and soil testing. We can see that the annual CO₂e emissions are low and almost in a similar range for the first 4 years and start to show higher value as the age increases across all the blocks.

5.2. Block-level hotspot identification

The annual carbon footprint of the farm blocks along with the plantation data detailed in Section 4.2 is used for hotspot identification. However, we do not use meteorological variables such as temperature and solar irradiation in identifying the hotspot blocks, even though these variables were used in the DNDC-based estimation of carbon footprint. This is because since the plantation blocks are in close proximity, these variables have similar values across the blocks and, therefore, fail to provide significant discriminatory insights.

To begin, we utilize the emission prediction module to assess the carbon footprint of palm blocks at the block level. We apply a range of well-established machine learning models, such as linear regression, decision tree, random forest, gradient boosting, and MLP regressor. Notably, the random forest regressor exhibits superior performance compared to the other models. The process of determining the optimal hyperparameters for the random forest models is conducted through thorough grid search experiments.

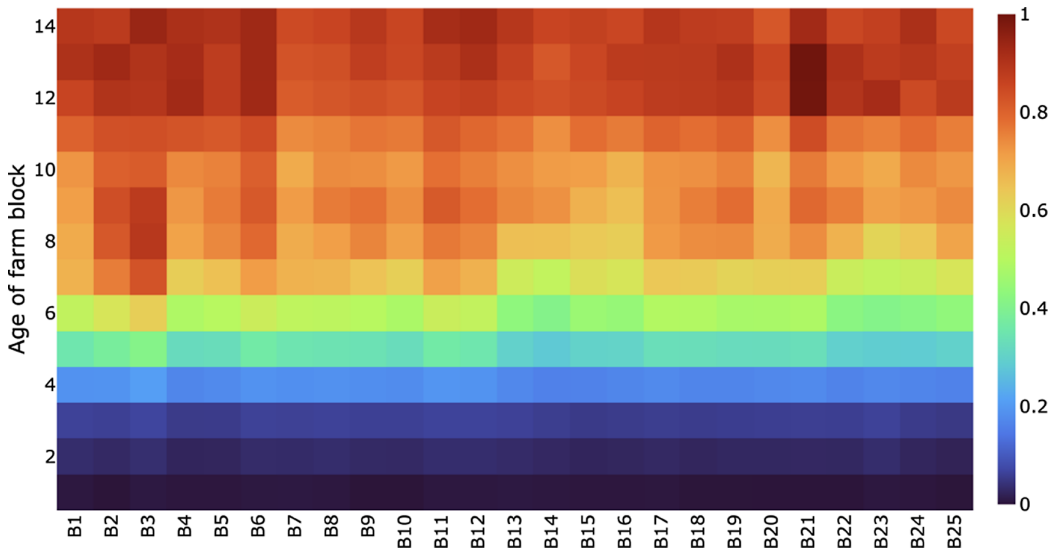


Figure 7. Normalized annual equivalent carbon emission (CO₂e) from farm blocks at different ages.

This involves methodically examining different hyperparameter values to identify the setup that maximizes performance, ensuring a well-informed and precise model selection.

The predicted carbon emissions serve as the expected emissions \mathbb{E} , to be used by the MDSS-based outlier detection module. We use the Gaussian scoring function (Speakman et al., 2016) as $\text{score}_{\text{bias}}$, which uses Gaussian distribution to model the log-likelihood ratio statistic defined in Eqn. (3.2). This serves as the statistical measure of divergence between subgroup observations and their expectations. Gaussian scoring function for a subgroup S is given by (Speakman et al., 2016):

$$\text{score}_{\text{bias}}(S) = \max_{q>1} \sum_{x_i \in d(S)} y_i \hat{\mu} \frac{(q-1)}{\hat{\sigma}^2} + \sum_{x_i \in d(S)} \hat{\mu}^2 \left(\frac{1-q^2}{2\hat{\sigma}^2} \right) \tag{5.1}$$

where, y_i are the observed values in the subset $d(S)$ belonging to subgroup S , $\hat{\mu}$, and $\hat{\sigma}^2$ are the expected mean and variance of the subgroup. We perform MDSS-based outlier detection, as outlined in 3.2.2 to obtain hotspot farm blocks, with $\text{score}_{\text{bias}}$ as defined in Eqn. (5.1). B2 and B3 are identified as hotspot blocks across multiple years. In Figure 8 the red region depicts the feature space of the anomalous subgroup S^* identified by MDSS. From this, we observe that blocks with very low or high fertilizer application, high manure application, and moderate yield are identified as hotspots. However, this does not provide explainable insights or help with identifying opportunities for reducing emissions. To address this, we will next analyze the hotspot blocks using explainability and what-if scenarios.

5.3. Hotspot explainability analysis

To understand the factors that dictate the variation in carbon footprints from palm plantations, we have used the SHAP value to identify the marginal contribution of all the relevant features. We begin with learning the prediction model with farming practices, soil properties, and weather-related parameters as input features for predicting the soil CO₂ and N₂O emissions independently. The individual emission regressor model will give us more insights into the dominating features that influence the respective emissions. Figures 9 and 10 show the holistic summary of feature importance for CO₂ and N₂O emission regressor models, respectively, using the SHAP value.

From Figure 9, we observe that for soil CO₂ emissions, the most dominating features are plantation age, the amount of nitrogen content in the manure and fertilizers, annual yield, annual precipitation, and

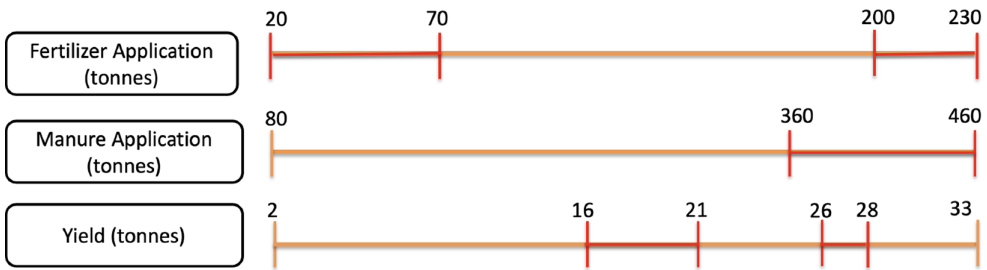


Figure 8. Anomalous subgroup identified by MDSS.

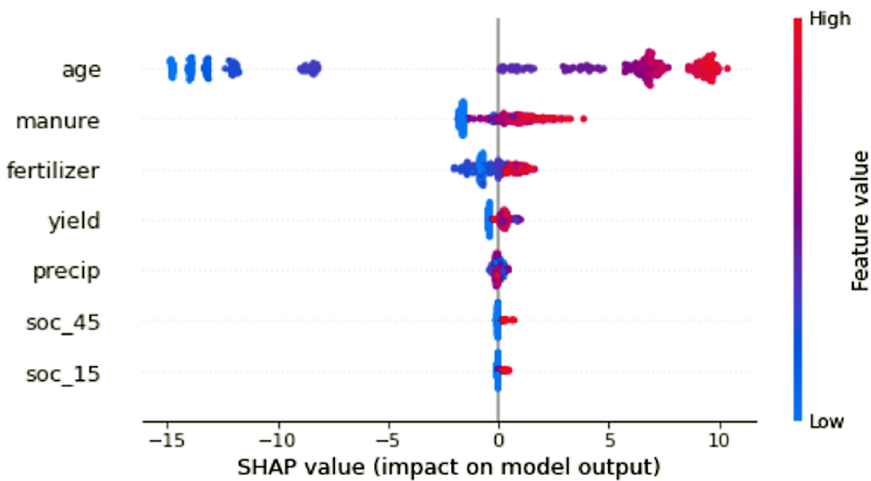


Figure 9. Global explainability for CO₂.

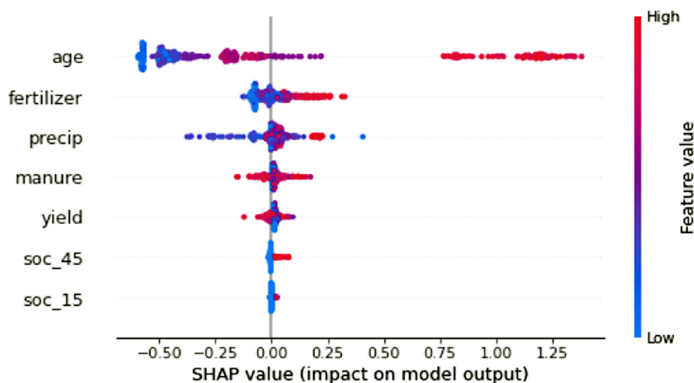


Figure 10. Global explainability for N₂O.

the initial soil organic carbon (SOC) at the depth of 15 cm–45 cm and 0–15 cm. Since the CO₂ emission is the product of soil decomposition it is clear that the above factors are influencing the rate of the decomposition process. The pruned palm fronds and the EFBs are the main sources of manure applications and have a high carbon-to-nitrogen ratio. Because of the high carbon content, the manure takes precedence over fertilizer in the feature importance plot. The amount of manure and fertilizers tends to increase with age and thus the increase in annual yield. For SOC at the depth of 0–15 cm and 15 cm–45 cm, we see a modest impact with a higher value leading to an increase in CO₂ while the lower value has

a neutral impact. This can be attributed to the capacity of soil to hold more carbon without getting saturated. The higher SOC content in the soil tends to have low holding capacity and thus more free carbon to participate in the decomposition process.

Similarly from Figure 10, we see that for N₂O emission, the dominating features are the same as those of CO₂ emission but the order of precedence and the impact are quite different. The nitrogen content in fertilizer takes precedence followed by annual precipitation and then the manure application. The N₂O emission is the product of the nitrification and denitrification process as shown in Figure 5. It is evident from the DNDC model workflow, that the presence of nitrate and ammonium ions with high precipitation tends to be a conducive environment for N₂O emission.

Since the dominating features of soil CO₂ and N₂O emissions are similar, it will be appropriate to investigate the feature importance of total carbon footprint (CO₂e). This will also capture the interdependency among soil CO₂ and N₂O emissions and help the enterprise focus on one metric for the overall carbon footprint reduction. Figure 11 shows the global interpretation of feature importance for the CO₂e emission regressor model using the Shapley value. The order of feature importance is the same as that of soil CO₂ and it shows that the proportion of soil CO₂ is larger than N₂O. Based on the features SHAP value, we infer that in general, the following factors lead to an increase in CO₂e—(i) high fertilizer and manure application results in higher emission. (ii) older blocks tend to have a higher carbon footprint associated with them. (iii) Soil organic carbon at depths 0–15 cm (SOC_45) and 15–45 cm (SOC_15) has a modest impact, with high soil organic carbon leading to an increase in carbon impact, while low SOC has a neutral impact.

We also perform an explainability analysis of the identified hotspot blocks to identify factors behind the low performance of the blocks. We use local explainability SHAP plots as mentioned in Section 3.3. The plot shows which factors contribute to the increase or decrease in CO₂e of the hotspot blocks when compared to the baseline of average carbon footprint. Figures 12 and 13 provide the plots for two hotspot blocks.

From Figure 12, we observe that high SOC_15 and SOC_45 content, high manure application result in an increase in emissions, while low fertilizer application lowers the carbon footprint. The block also has a

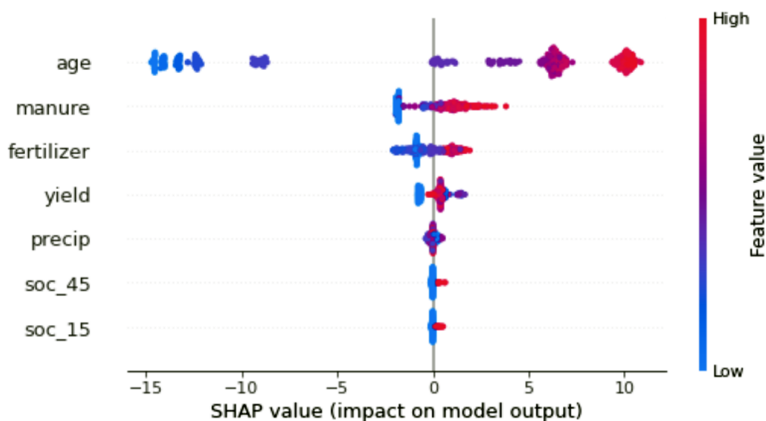


Figure 11. Global explainability for CO₂e emission.

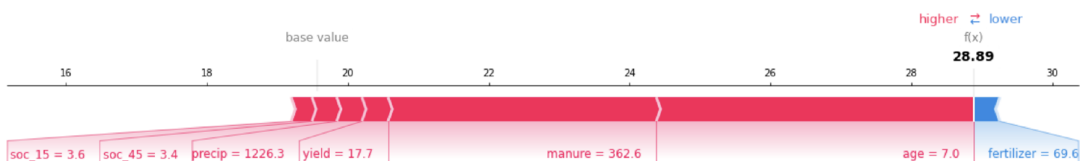


Figure 12. Palm plantation block B2 - local explainability.

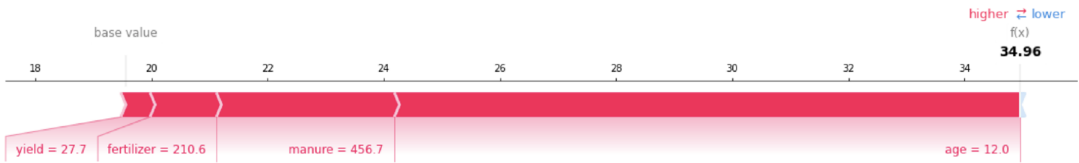


Figure 13. Palm plantation block B3 - local explainability.

moderate yield which could be due to low fertilizer application. From Figure 13, we can infer that very high fertilizer and manure applications are responsible for high emissions. The block has a good yield performance; however, it has a low carbon performance when compared to other blocks. The observations are in accordance with the insights drawn from Figure 11 earlier.

The two explainability plots share some common factors, we can observe that high fertilizer and high manure application negatively impact the carbon performance of the farm blocks.

5.4. Counterfactual recommendation

The recommendation Engine introduced in Section 3.4 is used to generate counterfactual recommendation plans for the hotspot blocks to reduce their carbon footprint. We focused on finding the set of intervenable features which can be perturbed to reduce the CO₂e emission without compromising on the yield produced by palm plantations. Figures 14 and 15 show the intervention plans generated by the engine for the two hotspot blocks. We observe that for both blocks, the fertilizer and manure application are identified as the intervenable areas. For B2, we generated the counterfactual query to produce a set of

fertilizer	manure	precip	age	soc_15	soc_45	yield	CO2e
69.56	362.58	1226.3	7.0	3.55	3.35	17.721434	28.891819

Diverse counterfactual set - CO₂e : [10,25]

fertilizer	manure	precip	age	soc_15	soc_45	yield	CO2e
63.3	248.2	-	-	-	-	17.5	23.27286577079987
58.0	266.5	-	-	-	-	18.4	24.36235165660021
63.5	241.7	-	-	-	-	17.1	22.794548305000127

Figure 14. B2 counterfactual recommendations.

fertilizer	manure	precip	age	soc_15	soc_45	yield	CO2e
210.58	456.69	1074.6	12.0	1.01	1.03	27.698867	34.865427

Diverse counterfactual set - CO₂e : [10,32]

fertilizer	manure	precip	age	soc_15	soc_45	yield	CO2e
213.3	326.9	-	-	-	-	27.8	31.083256241399727
194.1	368.6	-	-	-	-	28.6	31.1429671037998
192.9	333.6	-	-	-	-	28.5	30.29656883919958
216.5	321.8	-	-	-	-	26.5	31.29766918619998

Figure 15. B3 counterfactual recommendations.

three best diverse recommendations to restrict CO₂e emission within the range of 10 to 25 tonnes of CO₂e. From [Figure 14](#), we observe that reducing fertilizer application by around 10% and manure application by around 30% can help in reducing carbon emissions by roughly 20%, with negligible impact on the yield. Similarly, for B3, we generated the counterfactual query to produce the diverse set of four best recommendations without impacting the yield. From [Figure 15](#), we see that reducing manure application by 25%–30% can reduce the carbon emission by around 10% with negligible impact on the yield for the B3 palm plantation block.

6. Discussions

The escalating global GHG emissions from international trade demand enterprises to disclose and reduce emissions. Our EDA integrates emission computation, hotspot identification, explainability, and what-if analysis. Unlike qualitative methods or quantitative approaches lacking explainability, EDA combines rigor with qualitative insights. Application to the palm oil supply chain demonstrates its efficacy, positioning EDA as a strategic tool for enterprises committed to sustainable practices. The identification of agriculture-related hotspots underscores the need for targeted interventions, like optimizing fertilizer application, revealing EDA's applicability beyond agriculture. Versatility across industries highlights its potential for broader environmental challenges. Comparisons with previous research showcase EDA's uniqueness, offering industry-agnostic solutions for carbon footprint management. The proposed framework is generic, industry agnostic, and can be used across enterprises, emphasizing its significance in sustainable enterprise management.

7. Conclusion

In this paper, we presented a unified novel framework called EDA to accurately estimate the carbon footprint at the enterprise or process level across all types of asset classes, identify GHG hotspots and explain the factors that create the hotspots, and carry out what-if analysis to reduce the carbon footprint. The efficacy of the framework is demonstrated with palm oil enterprise data. Results presented in this paper indicate that the agriculture stage is the most carbon-intensive and blocks in which high fertilizer amounts were applied and low yields were obtained are the hotspots. This enabled customization of farming practices such as the right amount of fertilizer or manure for low-performing blocks which would not only improve their yield but also reduce their carbon footprint and sustainably ensure profitability.

8. Future works

In our future work, we plan to extend the EDA framework to address additional challenges and explore new opportunities for decarbonization. Specific areas of focus include:

- **Environmentally extended input–output (EEIO) based Scope 3 emissions:** Enhancing the framework to integrate and analyze Scope 3 emissions using the EEIO (Castellani et al., 2019) approach, providing a more scalable overview of an enterprise's environmental impact, especially when faced with a lack of physical activity data.
- **Diversification into other sectors:** Experimenting with the framework with data from other sectors such as the manufacturing industry, aviation industry, and real estate.
- **Stakeholder engagement and validation:** Collaborating with enterprises and stakeholders across various industries will be crucial for validating and refining the EDA framework. Real-world feedback and engagement will enhance the reliability and practicality of the framework in diverse business environments.
- **Long-term strategic recommendations:** While this framework provides short-term operational recommendations, in the future, we plan to extend this framework for long-term strategic recommendations to accelerate the net-zero decarbonization journey of enterprises.

These ongoing endeavors seek to refine and expand the EDA framework, ensuring its relevance and impact in diverse industries and contributing to the broader mission of sustainable decarbonization.

Author contribution. Data curation: H.S.; Writing – review & editing: H.S.; Conceptualization: J.H., R.G., S.G.; Methodology: J.H., M.P., A.J.; Software: J.H., M.P., A.J.; Writing – original draft: J.H., M.P., A.J.; Formal analysis: M.P., A.J.; Investigation: M.P., A.J.; Project administration: R.G., S.G.; Resources: R.G., S.G.; Supervision: R.G., S.G.; Visualization: A.J.

Competing interest. The authors declare none.

Data availability statement. The data belongs to a client and there is a non-disclosure agreement between IBM and the client to not share the data outside.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This work received no specific grant from any funding agency, commercial or not-for-profit sectors.

References

- Acquaye AA, Wiedmann T, Feng K, Crawford RH, Barrett J, Kuylenstierna J, Duffy AP, Koh SL and McQueen-Mason S (2011) Identification of ‘carbon hot-spots’ and quantification of GHG intensities in the biodiesel supply chain using hybrid LCA and structural path analysis. *Environmental Science & Technology* 45(6), 2471–2478.
- Arshad F, Tan YA and Yusoff S (2017) A cradle-to-gate study of GHG emissions from the transportation of palm oil, palm olein and palm stearin using the life cycle assessment approach. *Journal of Oil Palm Research* 29(1), 120–129.
- Babu YJ, Li C, Frolking S, Nayak DR and Adhya TK (2006) Field validation of DNDC model for methane and nitrous oxide emissions from rice-based production systems of India. *Nutrient Cycling in Agroecosystems* 74, 157–174.
- Bienge, K., Von Geibler, J., & Lettenmeier, M. (2010). Sustainability hot spot analysis: a streamlined life cycle assessment towards sustainable food chains. Building Sustainable Rural Futures: The Added Value of Systems Approaches in Times of Change and Uncertainty. 9th European IFSA Symposium, Vienna, Austria, 4-7 July 2010., 1822–1832.
- Black R, Cullen K, Fay B, Hale T, Lang J, Mahmood S and Smith S (2021) Taking stock: a global assessment of net zero targets. *Energy & Climate Intelligence Unit and Oxford Net Zero* 23.
- Castellani V, Beylot A and Sala S (2019) Environmental impacts of household consumption in Europe: comparing process-based LCA and environmentally extended input-output analysis. *Journal of Cleaner Production* 240, 117966.
- WNZ Challenge (2021) *The Supply Chain Opportunity*. Geneva, Switzerland: World Economic Forum.
- GEI (2016) Supply Chain Carbon and Energy Footprint. Retrieved June 23, 2022 from <https://www.globalefficiencyintel.com/supply-chain-carbon-and-energy-footprint>.
- Gilhespy SL, Anthony S, Cardenas L, Chadwick D, del Prado A, Li C, Misselbrook T, Rees RM, Salas W, Sanz-Cobena A, Smith P, Tilston EL, Topp CF, Vetter S and Yeluripati JB (2014) First 20 years of DNDC (denitrification decomposition): model evolution. *Ecological Modelling* 292, 51–62.
- Guo X, Broeze J, Groot JJ, Axmann H and Vollebregt M (2020) A worldwide hotspot analysis on food loss and waste, associated greenhouse gas emissions, and protein losses. *Sustainability* 12(18), 7488.
- Kuhndt M (2002) Hot spot analysis in practice—a case study focusing on MNC. Confidential Report/Project.
- Liedtke C, Baedeker C, Kolberg S, et al. (2010) Resource intensity in global food chains: the hot spot analysis. *British Food Journal*, 112(10), 1138–1159.
- Lundberg SM and Lee S-I (2017) A unified approach to interpreting model predictions. In GuyonI, LuxburgUV, BengioS, WallachH, FergusR, VishwanathanS and GarnettR (eds.), *Advances in Neural Information Processing Systems* 30. Red Hook, NY: Curran Associates Inc, pp. 4765–4774.
- Mothilal RK, Sharma A and Tan C (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617.
- Muhamad H, Sahid IB, Surif S, Ai TY and May CY (2012) A gate-to-gate case study of the life cycle assessment of an oil palm seedling. *Tropical Life Sciences Research* 23(1), 15.
- Myhre G, Shindell D, Bréon F-M, Collins W, Fuglestedt J, Huang J, Koch D, Lamarque J-F, Lee D, Mendoza B, Nakajima T, Robock A, Stephens G, Takemura T and Zhang H (2013) *Anthropogenic and Natural Radiative Forcing*. Cambridge, UK: Cambridge University Press, pp. 659–740.
- Myhre G, Shindell D, Bréon F-M, Collins W, Fuglestedt J, Huang J, Koch D, Lamarque J-F, Lee D, Mendoza B, et al. (2014). In Tignor K, Allen M, Boschung SK, Nauels J, Xia A, Bex Y and Midgley PM (eds.), IPCC, 2013: Climate Change 2013: The Physical Science Basis. *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. UK and New York, NY, USA: Cambridge University Press, 1535 pp.
- Neill DB, McFowland E and Zheng H (2013) Fast subset scan for multivariate event detection. *Statistics in Medicine* 32(13), 2185–2208.
- Norris CB, Norris GA and Aulisio D (2014) Efficient assessment of social hotspots in the supply chains of 100 product categories using the social hotspots database. *Sustainability*.

- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E** (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12, 2825–2830.
- Piringer G, Bauer A, Gronauer A, Saylor MK, Stampfel A and Iris K** (2016) Environmental hot spot analysis in agricultural lifecycle assessments—three case studies. *Journal of Central European Agriculture*.
- Protocol GG and Initiative GGP** (2004) *A Corporate Accounting and Reporting Standard*. World Resources Institute and World Business Council for Sustainable Development.
- Rodrigues TO, Caldeira-Pires A, Luz S and Frate CA** (2014) GHG balance of crude palm oil for biodiesel production in the northern region of Brazil. *Renewable Energy* 62, 516–521.
- Rudnicka A, Kalinowski TB and Wieteska G** (2017) Co2 hotspots identification in supply chains of different products. *Studia Ekonomiczne* 321, 01.
- Singh A, Mishra N, Ali SI, Shukla N and Shankar R** (2015) Cloud computing technology: reducing carbon footprint in beef supply chain. *International Journal of Production Economics* 164, 462–471.
- Speakman S, Somanchi S, McFowland E and Neill DB** (2016) Penalized fast subset scanning. *Journal of Computational and Graphical Statistics* 25(2), 382–404.
- Subramaniam V, Hashim Z, Loh SK and Astimar AA** (2020) Assessing water footprint for the oil palm supply chain—a cradle to gate study. *Agricultural Water Management* 237, 106184.
- Thomassen MA, Dalgaard R, Heijungs R and De Boer I** (2008) Attributional and consequential lca of milk production. *The International Journal of Life Cycle Assessment* 13(4), 339–349.
- Zhang Z and Neill DB** (2016) Identifying significant predictive bias in classifiers. Preprint, [arXiv:1611.08292](https://arxiv.org/abs/1611.08292).