Engineering and Technology Journal e-ISSN: 2456-3358

Volume 10 Issue 09 September-2025, Page No.-6743-6749

DOI: 10.47191/etj/v10i09.04, I.F. – 8.482

Analysis of the Environmental Impact of Food Production in Indonesia using Machine Learning Models Based on FAO Data

Oktaviana Maria Noni Imil[1], I Kadek Agus Wahyu Raharja[2*], Gde Wikan Pradnya Dana[3]

[1,2,3]Computer Engineering Program, Faculty of Engineering and Planning, Warmadewa University

ABSTRACT: Food production in Indonesia is a critical pillar of national food security but has significant environmental impacts, including greenhouse gas emissions, land use, and water consumption. This study analyzes the environmental impact of major food production in Indonesia using a Machine Learning (ML) approach based on data from the Food and Agriculture Organization (FAO). Data from FAO.csv, Food_Production.csv, and total_population_reform.csv were processed to identify relationships between food commodities and their ecological footprints. ML techniques, such as clustering and the XGBoost model, were employed to group commodities based on environmental impact and predict total emissions with an RMSE of 3.07 $MtCO_2e$. Results indicate that commodities like rice, palm oil, and root crops have significant environmental impacts, with rice contributing the highest emissions at 374.684 $MtCO_2e$ (1960–2013). This study provides data-driven strategies to support the sustainability of Indonesia's food sector, aligned with green technology principles, through visualizations of supply and emissions for the top 10 commodities.

KEYWORDS: Food production, environmental impact, Machine Learning, sustainability, FAO data

I. INTRODUCTION

Indonesia, as an agrarian nation with a population exceeding 270 million, faces significant challenges in ensuring food security while preserving environmental sustainability. Food production, such as rice, palm oil, and root crops, dominates the national agricultural sector but also contributes significantly to climate change through greenhouse gas (GHG) emissions, land degradation, and intensive water and land use [1] [2] [3]. Agriculture accounts for over 14% of national GHG emissions, with the food sector being a major contributor through activities like land preparation, fertilizer use, and transportation of harvested products [4].

Globally, food production contributes approximately 26% to total GHG emissions, with a significant proportion originating from developing countries like Indonesia [2]. Rice, a staple commodity in Indonesia, has a high carbon footprint due to wetland cultivation methods that produce methane, a greenhouse gas 25 times more potent than carbon dioxide [5]. Additionally, the expansion of palm oil plantations has led to widespread deforestation, increasing carbon emissions and reducing biodiversity [6]. These challenges are exacerbated by rising food demand due to population growth, projected to reach 300 million by 2030 [7].

Previous studies have highlighted the relationship between food production and environmental impacts, but most focus on global or regional scales, often overlooking Indonesia's specific context [2] [3]. Research in Indonesia is frequently limited to qualitative analyses or incomplete datasets, hindering the development of data-driven mitigation strategies [8]. With technological advancements, Machine Learning (ML) approaches offer solutions for analyzing complex datasets like those provided by FAO, enabling accurate predictions and identification of environmental impact patterns [9].

The use of FAO data in this study provides an advantage due to its extensive coverage, including historical food production, emissions, and resource use data from 1961 to 2013 [1]. This data enables rich longitudinal analysis to understand long-term trends and relationships between variables. However, challenges such as inconsistent formats and missing values require careful preprocessing, addressed in this study through data cleaning and transformation techniques [10].

Machine Learning approaches, particularly the XGBoost model and clustering, were chosen for their ability to handle large datasets with non-linear relationships [9]. XGBoost, a tree-based algorithm, has proven effective in predicting agricultural GHG emissions with high accuracy [11]. Meanwhile, clustering enables the grouping of commodities based on their environmental impact characteristics, providing insights into which commodities are most critical for mitigation [12]. This study integrates both techniques to deliver a comprehensive analysis.

The objectives of this study are to analyze the environmental impact of major food production in Indonesia, use ML models to predict emissions based on production attributes, and develop data-driven mitigation strategies. Focusing on key commodities like rice and palm oil, this study aims to provide policy recommendations that support the sustainability of the food sector, aligned with national emission reduction targets and the Sustainable Development Goals (SDGs) [13].

This study is expected to fill research gaps by providing an in-depth quantitative analysis of food production in Indonesia, an area underexplored with ML approaches. The findings are anticipated to assist policymakers, researchers, and stakeholders in designing environmentally friendly food production strategies, including the adoption of precision agriculture technologies and more efficient resource management [14].

II. LITERATURE REVIEW

The environmental impact of food production has been extensively studied, with a consensus that agriculture significantly contributes to global greenhouse gas emissions, land use changes, and water resource depletion. Poore and Nemecek (2018) estimated that food systems account for 26% of global GHG emissions, with crops like rice and oil palm being major contributors due to methane production and deforestation [2]. Vermeulen et al. (2012) further emphasized that food systems in developing countries face unique challenges due to intensive agricultural practices and limited technological adoption [3].

In the Indonesian context, rice production is a primary focus due to its role in food security and its high environmental footprint. Methane emissions from flooded rice paddies are a significant concern, as noted by the IPCC (2014), which highlights methane's potency as a greenhouse gas [5]. Studies by Yusuf and Warr (2019) indicate that rice cultivation in Indonesia contributes significantly to national emissions, exacerbated by traditional farming practices [8]. Palm oil production, another critical commodity, has been linked to deforestation and biodiversity loss, with Fitzherbert et al. (2008) documenting its impact on tropical ecosystems [6].

Machine Learning applications in environmental analysis have gained traction for their ability to handle complex datasets. Chen and Guestrin (2016) introduced XGBoost, a scalable tree-boosting algorithm effective for predictive tasks in agriculture, including emission forecasting [9]. Zhang et al. (2020) demonstrated its application in predicting agricultural emissions, achieving high accuracy in complex systems [11]. Clustering techniques, as reviewed by Jain et al. (1999), have been used to group agricultural products based on environmental impacts, aiding in targeted mitigation strategies [12].

Global studies, such as those by Tilman et al. (2011) and Godfray et al. (2010), underscore the need for sustainable intensification to meet rising food demand while minimizing environmental harm [15] [16]. Searchinger et al. (2019) proposed strategies like precision agriculture to reduce emissions, which are relevant to Indonesia's context [17]. Springmann et al. (2018) and Clark et al. (2020) further highlight the urgency of aligning food systems with climate targets to achieve the 1.5°C goal [18] [19].

In Indonesia, the environmental impacts of palm oil have been extensively studied. Austin et al. (2019) and Meijaard et al. (2020) discuss its role in deforestation and carbon emissions, emphasizing the need for sustainable practices [22] [23]. Tubiello et al. (2021) provide a global perspective on food system emissions, noting that developing countries like Indonesia face data gaps that ML can address [24]. Rosenzweig et al. (2020) advocate for adaptive agricultural practices to mitigate climate impacts, aligning with this study's objectives [25].

Despite these advancements, few studies integrate ML with comprehensive datasets like FAOSTAT to analyze Indonesia's food production impacts. The FAO's extensive data (1961–2013) offers a robust foundation for longitudinal analysis, yet challenges like missing values and format inconsistencies require sophisticated preprocessing [1] [10]. This study addresses these gaps by applying XGBoost and clustering to predict emissions and identify high-impact commodities, contributing to the growing body of literature on sustainable agriculture.

III. RESEARCH METHOD

A. Data Sources and Collection

Data were obtained from:

1. FAO.csv: Historical food production data (1961–2013).
2. Food_Production.csv: Environmental impact indicators such as GHG emissions ($MtCO_2e$), land use (ha/kg), and water use (m^3/kg).
3. total_population_reform.csv: Population data as an external control.

All datasets were sourced from FAOSTAT [1] and combined to form an analytical data foundation.

B. Data Preprocessing

The data preprocessing steps consisted of several stages. First, the dataset was filtered based on the column Area = Indonesia to ensure that only relevant national data were included in the analysis. Next, irrelevant attributes such as Area Code, Item Code, and Unit were removed to simplify the dataset. The data were then transformed from a wide format to a long format by introducing a year variable, which allowed for easier time-series analysis. Finally, duplicate records and rows with Supply = 0 were removed to improve data quality and consistency before further analysis. The implementation of these steps was carried out using the main Python code.

Table 1. Data preprocessing code

python

```
import pandas as pd  # Import pandas library for data manipulation and analysis

fao = pd.read_csv('FAO.csv')
# Read the dataset "FAO.csv" into a pandas DataFrame named 'fao'

fao_id = fao[fao['Area'] == 'Indonesia']
# Filter rows where the 'Area' column equals 'Indonesia'

fao_id = fao_id.drop(columns=['Area Code', 'Unit'])
# Drop unnecessary columns: 'Area Code' and 'Unit' to simplify the dataset

fao_id = fao_id.melt(id_vars=['Area', 'Item'], var_name='Year', value_name='Supply')
# Transform dataset from wide format to long format, keeping 'Area' and 'Item' as identifiers
# Columns representing years are melted into a single column 'Year',
# and their values are placed in a new column called 'Supply'

fao_id = fao_id[(fao_id['Supply'] != 0) & (~fao_id.duplicated())]
# Remove rows where 'Supply' equals 0 and drop any duplicate rows
```

C. Machine Learning Models

The main model applied in this study was the XGBoost Regressor [9], which was used to predict CO_2e emissions based on selected food production input features. The dataset was divided into two parts, with 90% allocated for training and 10% for testing, to ensure reliable model evaluation. Prior to training, the features were standardized using StandardScaler to normalize the scale of the variables and improve model performance. The model's accuracy was assessed using the Root Mean Square Error (RMSE), which measures the average magnitude of prediction errors and is defined as:

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

where $y_i$represents the actual observed values, $\hat{y}_i$ denotes the predicted values, and n is the total number of observations. A lower RMSE value indicates better predictive performance of the model. The implementation of these steps was carried out using the main Python code.

Table 2. Model training code

python

```
from sklearn.model_selection import train_test_split  # For splitting dataset into training and testing sets

from sklearn.preprocessing import StandardScaler  # For feature standardization (mean=0, variance=1)

from xgboost import XGBRegressor  # XGBoost regression model

from sklearn.metrics import mean_squared_error  # Metric to evaluate model performance (MSE / RMSE)

# Split dataset into training (90%) and testing (10%) sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1)

# Initialize StandardScaler to normalize features
scaler = StandardScaler()

# Fit the scaler on training data and transform it
X_train_scaled = scaler.fit_transform(X_train)

# Apply the same transformation to test data (using parameters learned from training set)
X_test_scaled = scaler.transform(X_test)

# Initialize XGBoost Regressor model
model = XGBRegressor()

# Train the model using scaled training data
model.fit(X_train_scaled, y_train)

# Predict CO₂e emissions on the scaled test data
y_pred = model.predict(X_test_scaled)

# Calculate Root Mean Square Error (RMSE) between actual and predicted values
rmse = mean_squared_error(y_test, y_pred, squared=False)
```

IV. RESULTS AND DISCUSSION

A. Data Exploration

The exploratory analysis revealed that rice recorded the highest total supply at 1,164,444 tons, followed by palm oil with 175,650 tons during the period 1961–2013 (Figure 1). The dominance of rice is consistent with its role as the staple food for the majority of Indonesia's population, making it not only economically important but also socially and culturally significant. Palm oil, on the other hand, although supplied in smaller quantities compared to rice, holds a strategic position as one of Indonesia's major export commodities.

The data cleaning process, which included the removal of irrelevant attributes, structural normalization, and data transformation, resulted in a dataset that was well-prepared for advanced analysis. This step was crucial to ensure the accuracy and reliability of the results, since inaccuracies or duplicated records could distort both the model outcomes and the emission estimations.

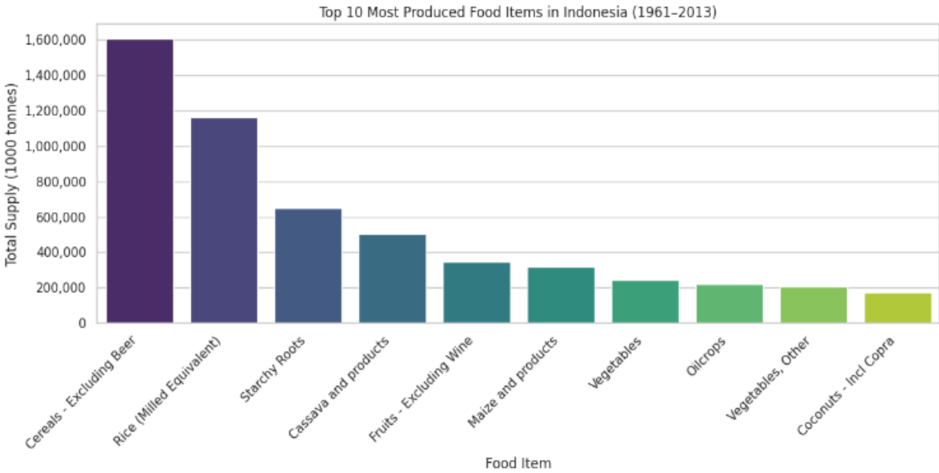

Figure 1: Commodities with the highest total supply (1961–2013).

B. Machine Learning Model Results

The XGBoost Regressor model achieved a relatively low error with an RMSE of 3.07 $MtCO_2e$, demonstrating its ability to predict emissions with strong reliability (Figure 2). Feature importance analysis highlighted land use as the most influential variable with a weight of 0.802, indicating that land-use change and intensity have the greatest effect on agricultural emissions. This finding aligns with existing literature, which emphasizes deforestation, agricultural expansion, and ecosystem degradation as primary drivers of greenhouse gas emissions.

Furthermore, the close alignment between predicted and actual values (Predicted vs. Actual curve) indicates that machine learning effectively captures non-linear relationships between agricultural production features and emissions. This confirms the potential of ML-based approaches to support data-driven environmental policy design.

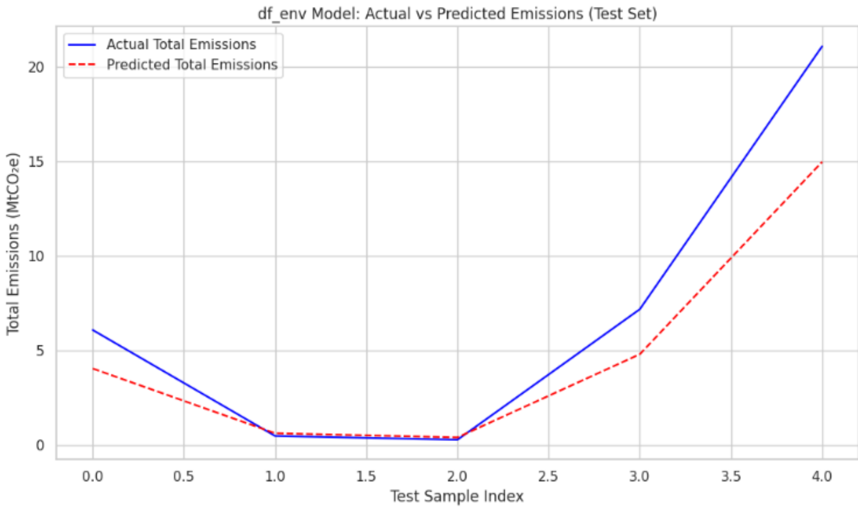

Figure 2: Predicted vs. actual emissions curve for the XGBoost model.

C. Environmental Impact of Commodities

The emission estimates for the period 1961–2013 show that rice contributed the largest share at 374.684 $MtCO_2e$, followed by palm oil (264.697 $MtCO_2e$) and other cereals (235.448 $MtCO_2e$) (Figure 3, Table 3). The dominance of rice reflects its massive domestic demand, but its cultivation is well known to release high levels of methane (CH_4) due to flooded paddy fields. Palm oil presents a dual challenge: while it is highly efficient in terms of oil yield per hectare, its expansion has often been associated with deforestation and biodiversity loss.

Table 3 further illustrates Indonesia's share of global agricultural emissions. Notably, coconuts (incl. copra) show a disproportionately high contribution of 25.23%, despite lower production volumes compared to rice or palm oil. This indicates that the emission intensity per unit production of coconut-related products is relatively high. Other commodities such as cassava (8.64%) and oilcrops (9.73%) also contribute significantly, suggesting that food system diversification comes with varying levels of emission trade-offs.

Table 3. Percentage contribution of Indonesia's emissions to global emissions

Item	Emissions Indonesia ($MtCO_2e$)	Emissions World ($MtCO_2e$)	Contribution (%)
Cassava	453,174	5,242,559	8.64
Cereals, Other	9,604,947	391,826,900	2.45
Coconuts – Incl. Copra	1,048,998	4,157,192	25.23
Fruits – Excluding Wine	2,982,881	112,922,900	2.64
Maize (Meal)	352,252	21,956,700	1.6
Oil crops	1,308,814	13,457,420	9.73
Other Vegetables	103,129	9,526,472	1.08
Rice	4,657,776	57,901,790	8.04
Roots & Tubers	3,884,799	135,635,400	2.86
Vegetables	1,457,627	144,404,700	1.01

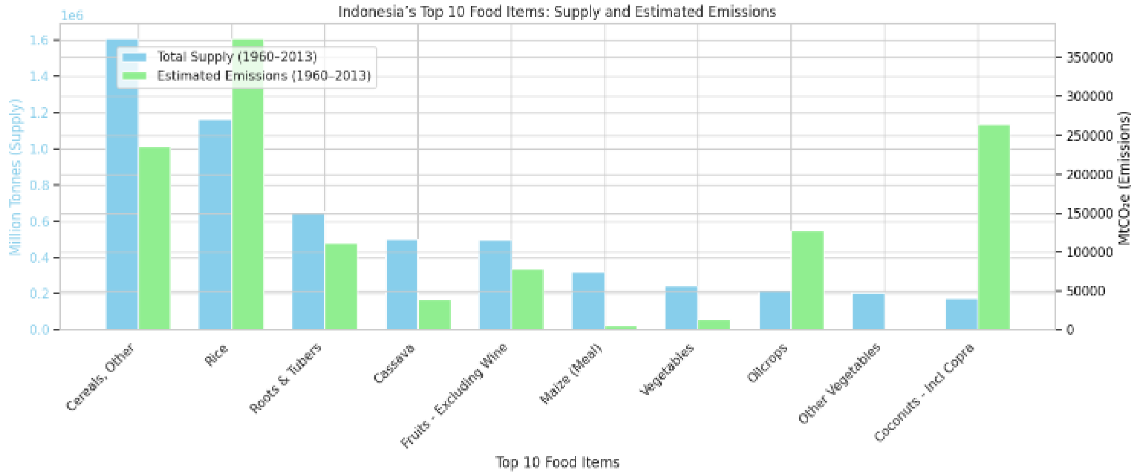

Figure 3: Emissions vs. supply for the top 10 commodities.

The visualization in Figure 3 highlights the strong relationship between supply and emissions, showing that commodities with the largest production volumes also tend to contribute the most to emissions. This pattern underlines a critical trade-off between food security and climate change mitigation. While intensifying production is necessary to meet growing food demand, it also carries the risk of increasing the national carbon footprint.

To address this challenge, sustainable agricultural strategies in Indonesia should focus on:

1. Improving land and water use efficiency, particularly in rice cultivation.
2. Adopting low-emission farming practices, such as alternate wetting and drying (AWD) in paddy systems.
3. Encouraging production diversification, to reduce dependence on high-emission commodities.
4. Leveraging machine learning for data-driven decision-making, to balance agricultural productivity with emission reduction targets.

V. CONCLUSIONS

This study demonstrates that machine learning approaches, particularly the XGBoost model, can effectively analyze and predict the environmental impacts of food production in Indonesia. The model achieved a low RMSE of 3.07 $MtCO_2e$, confirming its reliability in capturing the relationship between production features and emissions. The analysis identified rice and palm oil as the leading contributors to emissions, followed by other cereals, highlighting the need for targeted interventions in these sectors. To reduce environmental impacts, sustainable production strategies should prioritize efficient land and water management, the adoption of precision agriculture technologies, and the implementation of data-driven mitigation policies. These efforts are essential for balancing food security with climate change mitigation goals.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Food and Agriculture Organization (FAO) for providing access to the FAOSTAT database, which served as the primary data source for this study. The authors also extend their deepest appreciation to their course supervisor, for his invaluable guidance and support throughout the research process. Furthermore, the authors are grateful to the Computer Engineering Study Program, Faculty of Engineering and Planning, Warmadewa University, for providing the facilities and resources that made this study possible.

REFERENCES

1. FAO. (2020). FAOSTAT Database. Retrieved from http://www.fao.org/faostat/en/#data
2. Poore, J., & Nemecek, T. (2018). Reducing food’s environmental impacts through producers and consumers. *Science, 360*(6392), 987–992. https://doi.org/10.1126/science.aaq0216
3. Vermeulen, S. J., Campbell, B. M., & Ingram, J. S. I. (2012). Climate change and food systems. *Annual Review of Environment and Resources, 37*, 195–222.https://doi.org/10.1146/annurev-environ-020411-130608
4. Ministry of Environment and Forestry. (2018). *Indonesia Second Biennial Update Report.* Jakarta: Government of Indonesia. https://www.menlhk.go.id/site/single_post/2392
5. IPCC. (2014). *Climate Change 2014: Synthesis Report.* Intergovernmental Panel on Climate Change. https://www.ipcc.ch/report/ar5/syr/
6. Fitzherbert, E. B., et al. (2008). How will oil palm expansion affect biodiversity? *Trends in Ecology & Evolution, 23*(10), 538–545. https://doi.org/10.1016/j.tree.2008.06.012
7. Badan Pusat Statistik. (2020). *Proyeksi Penduduk Indonesia 2020–2050.* Jakarta: BPS. https://www.bps.go.id/publication/2020/09/30/proyeksi-penduduk-indonesia-2020-2050.html
8. Warr, P., & Yusuf, A. A. (2014). World food prices and poverty in Indonesia. *Australian Journal of Agricultural and Resource Economics, 58*(1), 1–21. https://doi.org/10.1111/1467-8489.12028
9. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939726
10. FAO. (2020). *The State of Food and Agriculture 2020.* Rome: Food and Agriculture Organization. https://www.fao.org/3/cb2867en/cb2867en.pdf
11. Zhang, Y., et al. (2020). Machine learning approaches for predicting agricultural emissions. *Environmental Modelling & Software, 126*, 104617. https://doi.org/10.1016/j.envsoft.2020.104617
12. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323. https://doi.org/10.1145/331499.331504
13. United Nations. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development.* New York: United Nations. https://www.un.org/sustainabledevelopment/sustainable-development-goals/
14. Wollenberg, E., et al. (2016). Reducing emissions from agriculture to meet the 2°C target. *Global Change Biology, 22*(12), 3859–3864. https://doi.org/10.1111/gcb.13340
15. Tilman, D., et al. (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences, 108*(50), 20260–20264. https://doi.org/10.1073/pnas.1116437108
16. Godfray, H. C. J., et al. (2010). Food security: The challenge of feeding 9 billion people. *Science, 327*(5967), 812–818. https://doi.org/10.1126/science.1185383
17. Searchinger, T., et al. (2019). *Creating a Sustainable Food Future.* World Resources Institute. https://www.wri.org/research/creating-sustainable-food-future
18. Springmann, M., et al. (2018). Options for keeping the food system within environmental limits. *Nature, 562*(7728), 519–525. https://doi.org/10.1038/s41586-018-0594-0
19. Clark, M. A., et al. (2020). Global food system emissions could preclude achieving the 1.5° and 2°C climate change targets. *Science, 370*(6517), 705–708. https://doi.org/10.1126/science.aba7357
20. Crippa, M., et al. (2021). Food systems are responsible for a third of global anthropogenic GHG emissions. *Nature Food, 2*, 198–209. https://doi.org/10.1038/s43016-021-00225-9
21. Carlson, K. M., et al. (2017). Greenhouse gas emissions intensity of global croplands. *Nature Communications, 8*, 13954. https://doi.org/10.1038/ncomms13954
22. Austin, K. G., et al. (2019). The environmental impacts of palm oil in context. *Nature Plants, 5*, 1251–1260. https://doi.org/10.1038/s41477-019-0548-2
23. Meijaard, E., et al. (2020). The environmental impacts of palm oil in context. *Nature Plants, 6*, 1418–1426. https://doi.org/10.1038/s41477-020-00813-w
24. Tubiello, F. N., et al. (2021). Greenhouse gas emissions from food systems: Building the evidence base. *Environmental Research Letters, 16*(6), 065007. https://doi.org/10.1088/1748-9326/ac018e

25. Rosenzweig, C., et al. (2020). Climate change and agriculture: Interactions, impacts, and adaptations. *Annual Review of Environment and Resources, 45*, 433–460. https://doi.org/10.1146/annurev-environ-012320-082542