Research article

Data-Driven Techniques for Identification of Factors Affecting Hydroponic Tomato: A Case Study in Hino City, Tokyo, Japan

HETTIGE SAMITHA LAKSHAN GUNASEKARA*
Graduate School of International Food and Agricultural Studies,
Tokyo University of Agriculture, Tokyo, Japan
Email: 13823001@nodai.ac.jp

RAMADHONA SAVILLE
Tokyo University of Agriculture, Tokyo, Japan

NINA N. SHIMOGUCHI
Tokyo University of Agriculture, Tokyo, Japan

KATSUMORI HATANAKA
Tokyo University of Agriculture, Tokyo, Japan

Received 31 December 2024 Accepted 27 May 2025 (*Corresponding Author)

Abstract Hydroponic greenhouses are a potential solution to the increasing demand for food and nutrition for the global human population. However, agriculture is influenced by complex relationships between multiple variables, particularly in regard to controlled environments, making it challenging for farmers to identify essential factors to create and manage optimal conditions for higher crop yields. Data-driven decision-making is an appealing solution for overcoming this challenge. The objective of this study was to use data science methods to identify the essential factors affecting a hydroponic tomato farm in Hino City, Tokyo, Japan. Specifically, this study identified the essential microclimatic and hydroponic factors that affect tomato yield. Further, this study compared the application of linear multiple regression and random forest regression models to identify the essential factors impacting the tomato harvest. Data sensors were installed in the greenhouse to obtain microclimatic and hydroponic data. Farm records, plant growth records, and tomato harvest data from three crop cycles (November 2021 to July 2024) were also collected. The moving average method was applied to smooth the data during preprocessing. The random forest regression model outperformed the linear regression model with a higher R^2 value of 0.9, whereas the linear model had a lower R^2 value of 0.31. Both models identified electrical conductivity supply, temperature, and the amount of water per plant as significant factors affecting tomato yield. Electrical conductivity showed a negative correlation, whereas temperature and the amount of water per plant showed a positive correlation, highlighting the importance of maintaining optimal levels for higher yields. This study provides practical insights into the essential yield-influencing factors and supports the implementation of customized management practices through data-driven decision-making, empowering smallholder hydroponic farmers to increase productivity.

Keywords linear multiple regression, machine learning, random forest regression, smart agriculture, yield optimization

INTRODUCTION

Agriculture continues to be the foundation of global food and nutritional security with more than 608 million family farms contributing significantly worldwide. Smallholder farmers account for 70–80 percent of global farmland and produce approximately 80 percent of the world's food by

value (FAO, 2021). In Japan, eighty percent of tomatoes are produced in smallholder greenhouses (MAFF 2022). Despite their pivotal role, smallholder farmers in Japan face multiple challenges, including unpredictable environmental conditions, resource limitations, labor shortages, and an aging farm worker population. These issues are particularly relevant in Japan, where smallholder farmers play a vital role in the agricultural sector. Therefore, the Japanese government has recognized the importance of supporting the agricultural sector by promoting smart agriculture (MAFF, 2024). Through subsidies and financial incentives, the government encourages farmers to adopt advanced technologies such as precision farming, data-driven decision-making, and automated systems (MAFF, 2023). These measures aim to mitigate labor shortages, enhance resource efficiency, and ensure the sustainability of smallholder farming operations. By integrating smart technologies, smallholder farmers can overcome traditional barriers and contribute to food security and economic sustainability in Japan. A promising approach for addressing these challenges is the adoption of smart hydroponic greenhouses. Smart hydroponic farming integrates hydroponic production techniques with sophisticated technologies such as Internet of Things (IoT) sensors, automated fertilizer delivery systems, and climate management to boost productivity and resource efficiency in a controlled setting (MAFF, 2024; Shareef et al., 2024). These systems offer a controlled environment for crop production in which nutrients are delivered through irrigation without soil, resulting in higher yields, better quality, and reduced resource use. This method is especially advantageous for smallholder farmers because it optimizes space, minimizes water consumption, and reduces dependency on traditional farming practices. However, managing hydroponic systems introduces complexities, requiring farmers to navigate the complex relationships between factors, such as temperature, humidity, CO_2 levels, solar radiation, and nutrient concentrations (Shareef et al., 2024). Even slight imbalances can significantly affect crop yields, underscoring the need for decision support tools.

Data-driven agriculture, a core component of smart agriculture, leverages technological advancements, IoT sensors, and data analytics to provide insights into farm management (Saiz-Rubio and Rovira-Más, 2020). In this type of agriculture, decisions are made by analyzing historical and real-time data by using data-driven techniques to help farmers identify critical variables, optimize growing conditions, and enhance productivity. Data science offers powerful yield prediction and optimization methodologies, ranging from traditional statistical models to advanced machine learning algorithms. The statistical method of linear multiple regression is effective in identifying linear relationships between variables (Breiman, 2001). However, this method has limitations in capturing complex nonlinear interactions. Machine learning regression methods can capture complex nonlinear interactions and have higher prediction power. Most machine learning algorithms focus solely on predictions. However, they cannot reveal how each variable impacts the predictive outcome or its importance (Carvalho et al., 2019). Random forest regression is a machine learning regression analysis method capable of yield prediction (Jeong et al., 2016). It also calculates feature or variable importance, providing insights into the contribution of each variable to the predicted outcome. This dual capability makes random forest regression particularly valuable in regard to understanding agricultural phenomena. Despite the potential of these data-driven techniques, their adoption in agriculture is limited to larger scale farms because smallholder farmers often face challenges such as a lack of awareness, technical expertise, and resources, leaving them unable to fully benefit from these advancements (Adereti et al., 2023). Therefore, addressing these barriers is crucial for democratizing access to smart-farming technologies and ensuring that smallholder hydroponic farmers can leverage data-driven methods to improve productivity and sustainability.

OBJECTIVE

The objective of this study was to identify the essential environmental and hydroponic variables that influence tomato production using data science methods for smallholder hydroponic farmers. Specifically, this study focused on identifying the microclimatic and hydroponic variables that significantly influence the tomato yield and analyzed the effectiveness of data-driven methods by

comparing the predictive power of linear multiple regression and random forest regression for the yield of hydroponic tomatoes.

METHODOLOGY

Study Design and Location

This study was conducted in a smallholder smart hydroponic greenhouse ("N Farm") in Hino City, Tokyo, Japan (Fig 1). While the N Farm greenhouse cultivates multiple tomato varieties, this study focused solely on the Momotaro tomato variety (660 plants grown inside a 237 m^2 area), as complete datasets (microclimatic, hydroponic, and farm management data) were available for this tomato variety. The research design involved collecting, preprocessing, and analyzing data from three consecutive crop cycles to identify key variables influencing tomato yield.

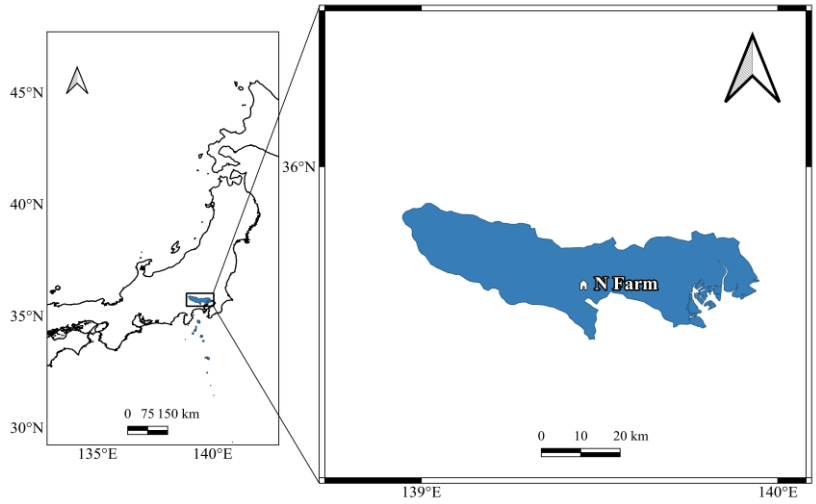

Fig. 1 The study location, Hino City, Tokyo Japan

Data Collection

Data was collected over three crop cycles from November 2021 to July 2024. The 1st cropping period was 2021-11-29 to 2022-06-27, the 2nd cropping period was 2022-10-30 to 2023-05-13, and the 3rd cropping period was 2023-10-16 to 2024-06-17. The study utilized a combination of microclimatic data (temperature [°C], CO_2 levels [parts per million - ppm], humidity [%], and cumulative sum of solar radiation [MJ/m^2]), hydroponic data (water supply per plant [ml], electrical conductivity (EC) [mS/cm] and pH value of the nutrition supply) (Fig 2), yield data (daily tomato harvest [kg]) and plant growth records (number of cultivating days, number of internodes) from the farm records (Fig 3) were used in this study.

Data Preprocessing

Handling missing data: Missing hydroponic data points were addressed using the forward-filling method, in which previous data values were replaced with missing values. This method ensures continuity without artificial fluctuations that can skew the analysis.
Moving average (MA): The MA technique was applied to reduce fluctuations and capture underlying trends. The window period was determined based on the average internode growth period of tomato cultivations.
Variance inflation factor value (VIF): A VIF analysis was conducted to detect multicollinearity among the independent variables in the regression models, as high multicollinearity can distort the

interpretation of the regression coefficients (Craney and Surles, 2002). The variables identified as having high multicollinearity (>10) were reviewed and adjusted to improve the performance and reliability of the regression models.

Data scaling: The dataset included variables with different units and scales. Before the analysis, the data were standardized using the Z-score method to ensure consistency and comparability, with a mean of zero and a standard deviation (SD) of one (D'Agostino et al., 2017). The standardized data were analyzed using linear multiple regression and random forest regression.

Fig. 2 Hydroponic data record by farmer

Fig. 3 Farmer demonstrates how plant growth data record

Data Analysis

Results from regression analysis: Two predictive models, linear multiple regression and random forest regression, were developed to analyze the relationships between the collected variables and tomato yield. Data aggregated from all three crop cycles were applied to these two models to explore the relationship between the independent variables, including environmental and hydroponic factors, and the dependent variable, the 11-day moving average (MA11) Momotaro tomato harvest.

Linear multiple regression: The linear multiple regression method was employed to explore the linear relationships between variables (Grégoire, 2014). The model estimates the impact of each independent variable on the tomato yield, providing coefficients that quantify their contribution to the predicted outcome.

Random forest regression: Random forest regression is an ensemble machine-learning algorithm that builds multiple decision trees and aggregates their predictions to deliver a more accurate and robust output for regression tasks (Breiman, 2001). By averaging the predictions from all individual decision trees, this method enhances overall accuracy and performance. This algorithm introduces randomness into the subset of features at each node for splitting, which reduces overfitting, enhances generalization, and effectively captures nonlinear relationships and complex interactions between variables, making it particularly useful for datasets with diverse features. Additionally, it provides insights into feature importance, helping identify the most influential factors in predicting the target variable (Gregorutti et al., 2017). This versatility, combined with its robustness to minimize noise and outliers, makes random forest regression a powerful tool for modeling in fields such as agriculture, where environmental and operational variables exhibit complex interdependence. The random forest regression model was trained on 70% of the dataset and tested on the remaining 30% of the data using a hyperparameter optimization grid search cross-validation to optimize the random forest regression model (Probst et al., 2019). For the random forest regression model, feature importance analysis was conducted to calculate the feature importance values and identify the most influential variables affecting tomato yield. After identifying the essential features, the Pearson correlation method was used to determine the effect of each variable on the 11-day Momotaro tomato harvest period.

Model Evaluation: The performance of the linear multiple regression and random forest regression models was evaluated using R^2 score, mean absolute error (MAE), and root mean squared error (RMSE) (Emmert-Streib and Dehmer, 2019). The R^2 score indicates the proportion of variance in the tomato yield that the model can explain, with higher values indicating better model performance. MAE measures the average magnitude of the prediction errors, indicating that the predictions are close to the actual values. The RMSE emphasizes the more significant errors by squaring them, indicating model accuracy with a bias toward more significant deviations.

RESULTS

The results provide insights into the key microclimatic hydroponic factors influencing tomato yield in the study's smallholder smart hydroponic greenhouse. Table 1 provides the descriptive results for all data variables used in this study for the three cropping periods as well as over the 1st to 3rd cropping periods. A higher average daily tomato harvest (15.4 ± 10.01 kg) was observed in the 1st cropping period, which declined gradually to 10.34 ± 6.34 kg by the 3rd cropping period.

Table 1 Descriptive statistics results for microclimatic and hydroponic data variables

	1st cropping period		2nd cropping period		3rd cropping period		1st to 3rd cropping period	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Harvest Data								
Daily Momotaro tomato harvest	15.44	10.01	11.50	8.04	10.34	6.34	12.96	8.52
Hydroponic Data								
Daily EC Supply	1.74	0.31	1.90	0.22	2.45	0.84	1.95	0.65
Daily pH Supply	6.64	0.40	6.33	0.41	5.77	0.77	6.41	2.25
Daily Total Water Per Plant	248.92	74.29	211.05	21.16	239.56	44.68	212.47	75.25
Microclimatic Data								
Daily Mean Temperature	18.25	3.37	17.35	1.61	19.33	4.00	19.05	3.71
Daily Humidity	85.68	8.67	87.70	6.60	81.61	10.65	83.60	9.45
Daily Mean CO_2	505.02	61.77	516.90	66.47	466.38	38.13	485.55	57.85
Daily Cumulative Sum of Solar Radiation	625.33	220.60	603.31	222.88	624.88	229.77	641.18	238.42
Growth Data								
Number of Cultivating Days	210	-	195	-	245	-	216.67	-
Number of Internodes	21	-	17	-	24	-	20.67	-

Also, when considering the hydroponic variables, EC increased from 1.74 ± 0.31 mS/cm in the 1st cropping period to 2.45 ± 0.84 mS/cm in the 3rd cropping period, suggesting that the farmer increased fertilization. The pH value measured each day in the hydroponic solution gradually decreased from 6.64 ± 0.40 in the 1st cropping period to 5.77 ± 0.77 in the 3rd cropping period. In addition, the water supply per plant had a higher value of 248.9 ± 74.29 ml in the 1st cropping period, which decreased to 211.05 ± 21.16 in the 2nd cropping period, but increased to 239.56 ± 44.68 ml in the 3rd cropping period. When considering the microclimatic variables, temperatures ranged from 17.35 ± 3.37 ℃ to 19.33 ± 4 ℃, averaging 19.05 ± 3.71 ℃ over the entire period. The mean daily humidity was highest at 87.7 ± 6.6% in the 2nd cropping period and dropped to 81.61 ± 10.65% during 3rd cropping period. The mean daily CO_2 levels recorded the highest value of 516.9 ± 66.47 ppm in the 2nd cropping period but decreased to 466.38 ± 38.13 ppm in the 3rd cropping period. The cumulative sum of solar radiation was similar for the 1st cropping period (625.33 ± 220.60 MJ/m^2) and the 3rd cropping period (624.88 ± 229.77 MJ/m^2) but decreased to 603.31 ± 222.88 MJ/m^2 in the 2nd cropping period. The average number of cultivation days for three crop cycles was 26.67 days. In the three crop cycles, the tomato plants had an average of 20.67 internodes, which was rounded to 21. Therefore, the average internode growth period was 10.34 days, which was rounded to 11 days. Figure 4 illustrates the daily Momotaro tomato harvest and the 11-day moving average (MA11) trend for the 3rd cropping period. The 3rd cropping period has

been selected for visualization purposes because it represents the most recent and complete dataset, allowing a clear illustration.

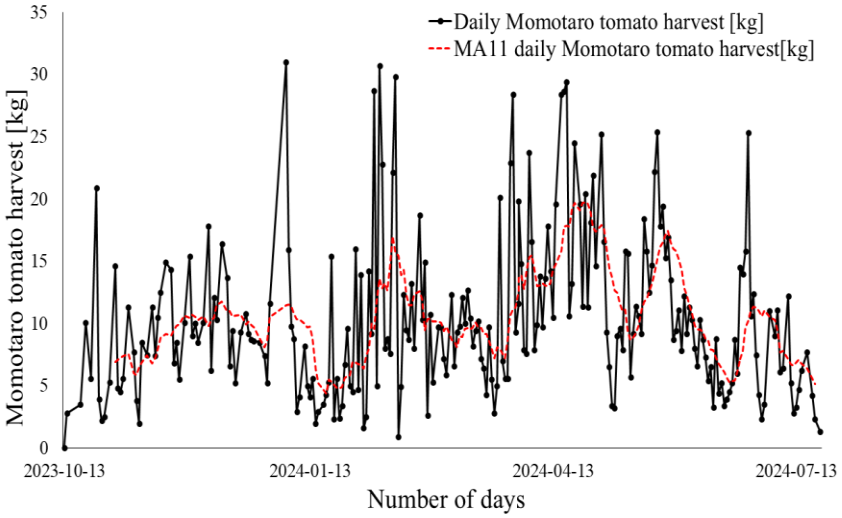

Fig. 4 Daily and MA11 days of Momotaro tomato harvest trend for 3rd cropping period

Regression Analysis Results

After preprocessing the dataset using the MA method, the dataset was analyzed using the VIF method to identify intercorrelated variables. The MA11 daily humidity had a high VIF value (>10). After removing the MA11 daily humidity variable, the remaining variables (which had VIF values below five) were standardized and used for the regression analyses.

Linear Multiple Regression

First, a linear regression model was developed to analyze the effects of microclimatic and hydroponic variables on tomato yield. The coefficient estimates and their significance levels are summarized in Table 2. The results showed that the model explained approximately 31% of the variability in the dependent variable, with an R^2 value of 0.31. Additionally, the low MAE of 0.67 and RMSE value of 0.81 highlight the predictive accuracy of the linear multiple regression model. The positively significant variables included the MA11 daily mean temperature, MA11 daily mean CO_2, and MA11 daily cumulative sum of solar radiation. Notably, it was seen that the MA11 daily EC supply was the only significant negative variable.

Table 2 Results from linear multiple regression analysis

Variables	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	0.00	0.03	0.00	1.00
MA11 Daily EC Supply	-0.37	0.04	-9.07	< 2e-16
MA11 Daily pH Supply	0.02	0.05	0.32	0.75
MA11 Daily Total Water Per Plant	0.06	0.03	1.87	0.06
MA11 Daily Mean Temperature	0.25	0.04	5.72	0.00
MA11 Daily Mean CO_2	0.09	0.04	2.41	0.02
MA11 Daily Cumulative Sum of Solar Radiation	0.34	0.04	8.14	0.00

Random Forest Regression

Second, a random forest regression model was developed to analyze the impact of microclimatic and hydroponic variables on tomato yield. The hyperparameter optimization grid search cross-validation including 100 trees ("n_estimators") with a maximum depth of 20, minimizing the split size of ("min_samples_split") and square root feature selection ("max_features"), which improved

the random forest regression model prediction performance offering its reliability for yield forecasting. The random forest regression model achieved an R^2 score of 0.9 on the test dataset, which explained 90% of the variability in the tomato harvest. Furthermore, the low MAE (2.63) and RMSE (1.63) values demonstrate higher prediction accuracies. As shown in Table 3, the random forest feature importance scores are reported alongside the correlation coefficients between variables and yield. The feature importance scores revealed that the most influential variables were the MA11 daily mean temperature value, MA11 daily EC supply, and MA11 daily total water per plant, which explained more than 63% of the variance affecting the MA11 daily Momotaro tomato harvest. Following the MA11 pH supply, the MA11 daily means CO_2 and MA11 daily cumulative sum of solar radiation were moderately important, meaning that they had a lower but still meaningful contribution to the model, with individual importance scores ranging between 0.11 and 0.15. The MA11 cumulative sum of solar radiation showed the lowest feature importance score of 0.11, and it exhibited the highest positive correlation (0.42) with the MA11 daily Momotaro tomato harvest. These results highlighted the essential environmental and hydroponic variables for hydroponic greenhouse productivity. After calculating the random forest regression, Pearson's correlation was used to examine the relationships between variables. The MA11 daily EC supply was negatively correlated with the MA11 daily Momotaro tomato harvest. In contrast, MA11 daily mean temperature and MA11 daily total water per plant were positively correlated with MA11daily Momotaro tomato harvest.

Table 3 Feature importance and correlation analysis from random forest regression

Variables	Feature Importance	Correlation
MA11 Daily Mean Temperature	0.22	0.26
MA11 Daily EC Supply	0.21	-0.28
MA11 Daily Total Water Per Plant	0.18	0.13
MA11 Daily pH Supply	0.15	0.04
MA11 Daily Mean CO_2	0.13	-0.10
MA11 Daily Cumulative Sum of Solar Radiation	0.11	0.42

DISCUSSION

The MA method with an 11-day window was used to smooth the dataset, with the window size determined based on the average internode growth period. An internode refers to the segment of a plant stem between two nodes, where branches and leaves develop during flower growth (Korol, 2022). In tomato plants, internodes are key indicators of growth and directly correlate with stages such as flowering and fruiting, which are critical for managing nutrient and water supply. Aligning the smoothing window with the internode growth period revealed biologically relevant patterns in the data. Subsequently, the prepared data were analyzed using the linear multiple regression and random forest regression methods.

When considering the linear multiple regression and random forest regression results, random forest regression had a higher model performance (R^2 score), explaining 90% of the phenomenon. In contrast, the linear multiple regression had a value of only 0.31, which explained 31% of the variance. This indicates that the random forest regression can better identify the nonlinear relationship than linear multiple regression methods. When considering the linear multiple regression and random forest regression, it was identified that MA11 days of tomato harvest, microclimatic variables of MA11 daily mean temperature, hydroponic variables of MA11 daily EC supply, and MA11 daily total water per plant are important by significant value in linear multiple regression and importance value in random forest regression. The MA11 daily mean temperature and MA11 daily total water per plant were positively correlated with the MA11 daily Momotaro tomato harvest. These findings are consistent with previous studies. Ullah et al. (2021) reported that reduced water supply under deficit irrigation conditions negatively impacts tomato yield in soilless systems, which supports this study's identification of water per plant as a significant

positive factor. John and Stephen (2024) and Van Ploeg and Heuvelink (2005) explained in their studies that suboptimal temperatures limit photosynthesis and fruit development, aligning with our findings that increased temperatures improve yield performance. The negative relationship between EC and tomato yield is also supported by Solis-Toapanta et al. (2020), who found that excessive EC concentrations delay ripening and increase susceptibility to blossom-end rot. This suggests that overfertilization may be occurring in the greenhouse, resulting in diminished productivity. Lee et al. (2017) further observed that nutrient uptake efficiency declines under high EC conditions, reinforcing the importance of maintaining a balanced nutrient supply. The MA11 daily cumulative sum of solar radiation showed the lowest feature of importance in the random forest model. However, it had the highest positive correlation with the MA11 daily Momotaro tomato harvest and was highly significant in the linear regression model. This suggests that solar radiation plays a biologically important, yet partly indirect, role in tomato yield, likely through its influence on photosynthesis, fruit development, and the regulation of tomato growth (Teixeira, 2020). In contrast, variables such as MA11 days mean temperature, MA11 daily total water per plant, and MA11 daily EC supply were consistently identified as primary drivers of yield by both models. The comparatively negative lower feature importance for solar radiation may result from shared variance with other variables, which often dominates in model splits in tree-based algorithms (Gregorutti et al., 2017). Nevertheless, the strong correlation and statistical significance of solar radiation explain that it should not be overlooked in greenhouse management. Therefore, model-derived feature importance should always be interpreted in conjunction with statistical evidence, biological mechanisms, and expert agronomic knowledge to avoid dismissing variables that contribute meaningfully to crop productivity.

While these findings support prior research, this study also extends the literature in several keyways. First, it is one of the few to apply both linear regression and random forest regression to small-scale hydroponic greenhouse data, enabling comparative analysis of model performance and variable importance. Second, the use of an internode-aligned moving average window provides a novel approach that links physiological plant growth stages to data preprocessing, thereby enhancing the biological relevance of the analysis. Lastly, the study focuses specifically on smallholder hydroponic production in Japan, a context often overlooked in existing data-driven agriculture studies, such as Jeong et al. (2016), which primarily address larger scale farm systems.

According to the descriptive results, the hydroponic greenhouse had an average temperature of 19.05 ± 3.71 ℃ while the optimal temperature range for tomatoes is approximately 25 °C to 30 °C (John and Stephen, 2024). Increasing the temperature is important for increasing tomato yield productivity, because suboptimal temperatures can decrease the rate of photosynthesis, slow plant metabolic processes, and negatively impact fruit development and ripening. Therefore, the farmer needs to take the necessary actions to increase their temperature levels. In addition, Ullah et al. (2021) mentioned in their study how a water deficit can result in a reduced tomato harvest. Consistent with this finding, our results suggest that ensuring an adequate water supply is essential for maximizing productivity. The MA11 days of EC negatively affected the MA11 days of the tomato harvest. Therefore, it can be expected that the farmer has been overfertilizing and that higher EC values would reduce crop productivity. Therefore, based on the findings, consistent temperature controls must provide favorable conditions for photosynthesis, nutrient uptake, and overall plant health. As for the hydroponic variance variables, providing balanced nutrient management to mitigate the potential stress caused by over-fertilization and maintaining precise water management is essential to avoid over-irrigation, which can adversely affect the yield. Although the MA11 daily cumulative sum of solar radiation exhibited lower feature importance in the random forest model, it showed the highest positive correlation with tomato yield and was statistically significant in the linear multiple regression analysis. This indicates that solar radiation plays a meaningful, though partially indirect, role in yield performance. Therefore, the farmer should not overlook monitoring solar radiation levels and implement strategies to enhance light exposure, such as using reflective surfaces or installing supplementary lighting during periods of low radiation, which can contribute to improved crop yield.

CONCLUSION

This study aimed to identify the key microclimatic and hydroponic factors influencing tomato yield in a hydroponic greenhouse system and evaluate the predictive performance of data-driven models. Using environmental and hydroponic data from three crop cycles (2021-2024) at N Farm in Hino City, Tokyo, this study identified that MA11 daily mean temperature, MA11 daily EC supply, and MA11 daily total water per plant are the most important factors for the MA11 Momotaro tomato yield by using linear multiple regression and random forest regression models. Temperature and water supply showed positive correlations with yield. On the other hand, EC showed a negative effect, indicating yield penalties of over-fertilization. Although the MA11 daily cumulative sum of solar radiation received only lower feature importance in the random forest model, it had the highest positive correlation. Also, it was highly significant in the linear multiple regression, suggesting that it is likely due to overlapping effects with other variables. Despite this, solar radiation remains important and should be managed in a greenhouse alongside other factors. Random forest regression captured these complex nonlinear relations more effectively ($R^2 = 0.90$) than the linear multiple regression model ($R^2 = 0.31$). In addition to these results, this study offers several significant contributions. First, it integrates environmental and hydroponic data with a biologically informed 11-day moving average aligned with the tomato internode growth period, enhancing the physiological relevance of the analysis. Second, this study compared statistical regression with machine learning in a hydroponic setting, demonstrating that machine learning models can deliver both high accuracy and the importance of interpretable variables at this scale. Third, concentrating on an operational Japanese hydroponic greenhouse fills a practical gap in the literature, as data-driven techniques are feasible and valuable for smallholder farmers. These findings and methods create a replicable, data-driven framework that delivers practical guidance for temperature control, nutrient balance, and irrigation scheduling. Therefore, this study implements customized management practices through data-driven decision-making, empowering smallholder hydroponic farmers to increase productivity.

The limitations of this study include identifying only the essential variables and their effects on tomato yields using historical data. Therefore, future studies should focus on identifying the optimal conditions for each identified variable, the applicability of real-time data integration, and broader applications across diverse smallholder hydroponic farming systems to enhance the impact of data-driven agriculture.

ACKNOWLEDGEMENTS

The authors sincerely thank N Farm for their invaluable support and collaboration throughout this study. The provision of the hydroponic greenhouse facilities and assistance with data collection were instrumental in the successful completion of this research.

REFERENCES

Adereti, D.T., Gardezi, M., Wang, T. and McMaine, J. 2023. Understanding farmers' engagement and barriers to machine learning-based intelligent agricultural decision support systems. Agronomy Journal, 116 (3), 1237-1249, Retrieved from DOI https://doi.org/10.1002/agj2.21358

Breiman, L. 2001. Random forests. Machine Learning, 45 (1), 5-32, Retrieved from DOI https://doi.org /10.1023/a:1010933404324

Carvalho, D.V., Pereira, E.M. and Cardoso, J.S. 2019. Machine learning interpretability: A survey on methods and metrics. Electronics, 8 (8), 832, Retrieved from DOI https://doi.org/10.3390/electronics 8080832

Craney, T.A. and Surles, J.G. 2002. Model-dependent variance inflation factor cutoff values. Quality Engineering, 14 (3), 391-403, Retrieved from DOI https://doi.org/10.1081/qen-120001878

D'Agostino, M., Dardanoni, V. and Ricci, R.G. 2017. How to standardize (if you must). Scientometrics, 113 (2), 825-843, Retrieved from DOI https://doi.org/10.1007/s11192-017-2495-7

Emmert-Streib, F. and Dehmer, M. 2019. Evaluation of regression models, Model assessment, model selection and generalization error. Machine Learning and Knowledge Extraction, 1 (1), 521-551, Retrieved from DOI https://doi.org/10.3390/make1010032

FAO. 2021. Small family farmers produce a third of the world's food. Newsroom, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy, Retrieved from DOI https://doi.org/10.2172/1329289

Grégoire, G. 2014. Multiple linear regression. European Astronomical Society Publications Series, 66, 45-72, Retrieved from DOI https://doi.org/10.1051/eas/1466005

Gregorutti, B., Michel, B. and Saint-Pierre, P. 2017. Correlation and variable importance in random forests. Statistical Computing, 27, 659-678, Retrieved from DOI https://doi.org/10.1007/s11222-016-9646-1

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K., Gerber, J.S., Reddy, V.R. and Kim, S. 2016. Random forests for global and regional crop yield predictions. PLoS ONE, 11 (6), e0156571, Retrieved from DOI https://doi.org/10.1371/journal.pone.0156571

John, A.A. and Stephen, R. 2024. Adaptation and mitigation of high temperature stress in tomato. International Journal of Environment and Climate Change, 14 (6), 322-331, Retrieved from DOI https://doi.org/10.9734/ijecc/2024/v14i64231

Korol, V.G. 2022. Growth of internodes and branching of a tomato plant. Vegetable Crops of Russia, 2, 15-19, Retrieved from DOI https://doi.org/10.18619/2072-9146-2022-2-15-19

Lee, J.Y., Rahman, A., Azam, H., Kim, H.S. and Kwon, M.J. 2017. Characterizing nutrient uptake kinetics for efficient crop production during Solanum lycopersicum var. cerasiforme Alef. growth in a closed indoor hydroponic system. PLoS ONE, 12 (5), e0177041, Retrieved from DOI https://doi.org/10.1371/journal.pone.0177041

MAFF. 2022. Crop status survey (vegetables). Ministry of Agriculture, Forestry and Fisheries (MAFF), Japan, Retrieved from URL https://www.maff.go.jp/j/tokei/kouhyou/sakumotu/sakkyou_yasai/

MAFF. 2023. Next-generation agricultural support service. Ministry of Agriculture, Forestry and Fisheries (MAFF), Japan, Retrieved from URL https://www.maff.go.jp/j/kanbo/smart/attach/pdf/index-171.pdf

MAFF. 2024. Regarding the situation surrounding smart agriculture. Ministry of Agriculture, Forestry and Fisheries (MAFF), Japan, Retrieved from URL https://www.maff.go.jp/hokuriku/seisan/smart/attach/pdf/index-12.pdf

Probst, P., Wright, M.N. and Boulesteix, A. 2019. Hyperparameters and tuning strategies for the forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9 (3), Retrieved from DOI https://doi.org/10.1002/widm.1301

Saiz-Rubio, V. and Rovira-Más, F. 2020. From smart farming towards agriculture 5.0: A review on crop data management. Agronomy, 10 (2), 207, Retrieved from DOI https://doi.org/10.3390/agronomy10020207

Shareef, U., Rehman, A.U. and Ahmad, R. 2024. A systematic literature review on parameters optimization for smart hydroponic systems. AI, 5 (3), 1517-1533, Retrieved from DOI https://doi.org/10.3390/ai5030073

Solis-Toapanta, E., Fisher, P.R. and Gómez, C. 2020. Effects of nutrient solution management and environment on tomato in small-scale hydroponics. HortTechnology, 30 (6), 697-705, Retrieved from DOI https://doi.org/10.21273/horttech04685-20

Teixeira, R.T. 2020. Distinct responses to light in plants. Plants, 9 (7), 894, Retrieved from DOI https://doi.org/10.3390/plants9070894

Ullah, I., Mao, H., Rasool, G., Gao, H., Javed, Q., Sarwar, A. and Khan, M.I. 2021. Effect of deficit irrigation and reduced N fertilization on plant growth, root morphology and water use efficiency of tomato grown in soilless culture. Agronomy, 11 (2), 228, Retrieved from DOI https://doi.org/10.3390/agronomy11020228

Van Ploeg, D. and Heuvelink, E. 2005. Influence of sub-optimal temperature on tomato growth and yield: A review. The Journal of Horticultural Science and Biotechnology, 80 (6), 652-659, Retrieved from DOI https://doi.org/10.1080/14620316.2005.11511994