Machine learning-based modeling of net ecosystem exchange using numerical weather prediction data and satellite images

Nari Kim[1], Jaeil Cho[2], and Yangwon Lee[3#]

[1]Research Institute for Geomatics, Pukyong National University, 45 Yongso-ro, Nam-gu, Busan, South Korea

[2]Department of Applied plantScience, Chonnam Nationa University, 77 Yongbong-ro, Buk-gu, Gwangju, South Korea

[3]Department of Spatial Information Engineering, Pukyong National University, 45 Yongso-ro, Nam-gu, Busan, South Korea

[#]Corresponding author: modconfi@pknu.ac.kr

ABSTRACT

Recently, the increasing severity of climate change attributable to global warming has emphasized the imperative of carbon absorption to mitigate greenhouse gas emissions. The use of the carbon sink based on the carbon absorption and storage functions of forests is suggested as an effective alternative for domestic greenhouse gas reduction. Additionally, agricultural land cover comprises approximately 38% of the Earth's surface, underscoring the importance of comprehensively understanding the carbon cycle within not only forests but also agricultural landscapes. This significance arises from the fact that agricultural land locally amplifies seasonal variations in carbon dioxide by approximately 25% compared to vegetated areas. Consequently, a comprehensive understanding of both forest and agricultural land carbon cycles is imperative, necessitating quantitative analysis of carbon uptake in agricultural settings. Thus, this study aims to construct a machine learning-based model for estimating the net ecosystem exchange (NEE) of rice paddies in South Korea using ground flux data, meteorological variables, and satellite images. Through quantitative assessment, the NEE was determined, with a mean absolute error of 1.387, root mean square error of 2.203, and correlation coefficient of 0.872. Notably, observed NEE values demonstrating extremes in magnitude were associated with calculation errors, reflecting tendencies of both underestimation and overestimation. This phenomenon is likely attributed to the study's reliance on a limited dataset and the inherent challenges of training models across a broad spectrum of observations. To enhance calculation accuracy, future endeavors should focus on accumulating a more extensive repository of NEE flux observations and leveraging high-resolution satellite imagery and meteorological datasets for refining machine learning-based models.

Keywords: net ecosystem exchange; machine learning; satellite images

1. Introduction

In recent years, the increasing severity of climate change attributable to global warming has emphasized the imperative of carbon absorption to mitigate greenhouse gas emissions. The use of the carbon sink based on the carbon absorption and storage functions of forests is suggested as an effective alternative for domestic greenhouse gas reduction. Additionally, agricultural land cover comprises approximately 38% of the Earth's surface, underscoring the importance of comprehensively understanding the carbon cycle within not only forests but also agricultural landscapes. This significance arises from the fact that agricultural land locally amplifies seasonal variations in carbon dioxide by approximately 25% compared to vegetated areas (Satio et. 2005; Gray et al. 2014; Zeng et al. 2014). Consequently, a comprehensive understanding of both forest and agricultural land carbon cycles is imperative, necessitating quantitative analysis of carbon uptake in agricultural settings. Thus, this study aims to construct a machine learning-based model for estimating the net ecosystem exchange (NEE) of rice paddies in South Korea using ground flux data, meteorological variables, and satellite images.

2. Data and Method

2.1. Data

To estimate NEE, ground flux tower data and numerical weather prediction data were used.

The study was conducted by collecting data from flux towers installed in rice paddy in Naju (35.0275N, 126.8208E) and Cheorwon (38.2013N, 127.2506E), South Korea (Fig.1).

Figure 1. Study Area: (a) Naju Flux tower and (b) Cheorwon Flux tower

As satellite images, normalized difference vegetation index (NDVI), which represent vegetation growth and vitality, leaf area index (LAI) and fraction of photosynthetically active radiation (FPAR) from the Moderate Resolution Imaging Spectroradiometer (MODIS) were used.

As meteorological data, data from the Local Data Assimilation and Prediction System (LDAPS), a local forecast model operated by the Korea Meteorological Administration, were used. In this study, a total of 11 variables were used: shortwave radiation (SWR), longwave radiation (LWR), sensible heat flux (H), soil heat flux (G), latent heat flux (LE), wind speed (WS), air temperature (Ta), surface temperature (Ts), soil temperature (Tsoil10), relative humidity (RH), 0~10cm soil moisture (SM).

A total of 14 variables from satellite images and meteorological data were used to construct match-up dataset for the Naju and Cheorwon sites (Fig.2). Table 1 summarizes the dataset used in this study, and Fig. 2 indicates the process for constructing the machine learning-based modeling of NEE.

Table 1. The dataset used in this study

Data Source	Variables	Spatial resolution	Temporal resolution
Flux tower	Net Ecosystem Exchange (NEE) (Naju)	Point	30 minutes
	Net Ecosystem Exchange (NEE) (Cheorwon)	Point	Daily
VIIRS	Normalized Difference Vegetation Index (NDVI)	1 km	8days
GK2A AMI	Normalized Difference Vegetation Index (NDVI)	2 km	Daily
MODIS	Fraction of Photosynthetically Active Radiation (FPAR)	500 m	8 days
	Leaf Area Index (LAI)	500 m	8 days
LDAPS	Shortwave Radiation (SWR)	1.5 km	3 hours (00, 03, 06, 09, 12, 15, 18, 21 UTC)
	Longwave Radiation (LWR)		
	Sensible Heat Flux (H)		
	Soil Heat Flux (G)		
	Latent Heat Flux (LE)		
	Wind Speed (WS)		
	Air Temperature (Ta)		
	Surface Temperature (Ts)		
	Soil Temperature (Tsoil10)		
	Relative Humidity (RH)		
	Soil Moisture (SM)		

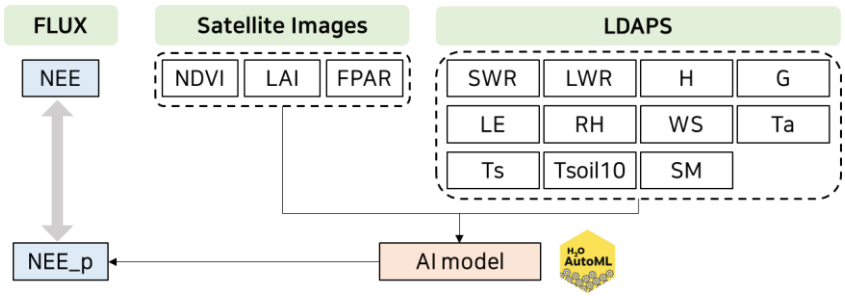

Figure 2. Process for constructing the machine learning-based modeling of net ecosystem exchange

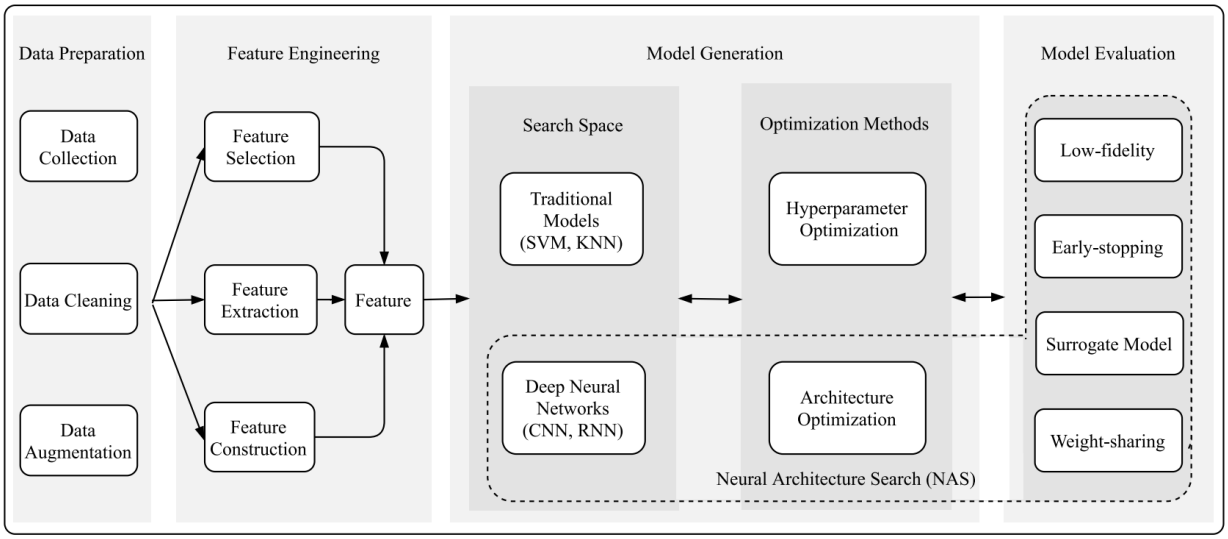

Figure 3. An overview of AutoML pipeline covering with 4 sections (Data Prepariation, Feature Engineering, Model generation and Model evaluation) (He et al., 2021)

2.2. Methods

Automated Machine Learning (AutoML) represents a computational framework designed to streamline pivotal tasks within machine learning workflows. These tasks encompass data extraction, model training, hyperparameter optimization, and exploration of neural network architectures (Jin et al., 2019). This automated process executes machine learning algorithms including Distributed Random Forest (RF), Generalized Linear Model (GLM), Extreme Gradient Boosting (XGBoost), Gradient Boosting Machine (GBM), and Deep Neural Network (DNN) using diverse initial parameter configurations. Subsequently, it discerns the optimal performing model by assembling an ensemble of the foremost N models (LeDel and Poirier, 2020). It autonomously manages tasks, such as algorithm selection and model tuning, which developers conventionally undertake manually and iteratively to enhance model performance. This automated approach culminates in efficiently optimized outcomes, as illustrated in Fig. 3. In this study, modeling was conducted utilizing the h2o library within the R programming environment. A cap of 20 models was imposed, and the assessment of model performance was conducted through 5-fold cross-validation.

3. Results

For the construction of the Net Ecological Exchange (NEE) model, observational data spanning from 2020 to 2021 for the Naju flux tower and from 2015 to 2018 for the Cheorwon flux tower were gathered. Additionally, satellite image data and LDAPS hydrometeorological data were employed as inputs for model development. Utilizing the gathered data, we constructed three sets of matchups: one for Naju, another for Cheorwon, and a third combining data from both locations. Subsequently, AutoML modeling was conducted independently for each of these matchups. AutoML modeling was executed for 221 cases in Naju, 1272 cases in Cheorwon, and 1493 cases in the combined dataset of Naju and Cheorwon. A total of 20 models were built for each dataset employing 5-fold cross-validation methodology. Additionally, two ensemble models were generated from the outcomes. After comparing the performance of the 22 models constructed for each matchup, it was determined that the Gradient Boosting Machine (GBM) model exhibited the highest performance across all three datasets. Scatterplots and accuracy statistics for each model are illustrated in Figure 4 and Table 1, respectively.

For Naju, the Mean Absolute Error (MAE) was determined to be 2.333 μmolCO2 m-2, the Root Mean Square Error (RMSE) was 3.678 μmolCO2 m-2, and the correlation coefficient was 0.604. Examination of the scatter plot results (Fig. 4(a)) reveals significant calculation errors particularly evident for both large and small values of NEE. For Cheorwon, the MAE was recorded at 1.139 μmolCO2 m-2, while the RMSE amounted to 1.707 μmolCO2 m-2. The CC was 0.924, underscores the model's high accuracy. Moreover, the scatterplot result (Fig. 4(b)) visually demonstrates a close distribution of observed and modeled values, aligning closely with a one-to-one line. Upon combining the data from Naju and Cheorwon, the calculated MAE was 1.387 μmolCO2 m-2, with a RMSE of 2.203 μmolCO2 m-2. The CC for this combined dataset was determined to be 0.872. From the scatterplot results (Fig. 4(c)), it is evident that calculation errors tend to be substantial, primarily attributed to the model's tendency to underestimate very large NEE values and overestimate very small observations. However, the errors are generally distributed in close proximity to the one-to-one line, indicating reasonable agreement between observed and modeled values across the spectrum. The presence of very large or small values of NEE can likely be attributed to the limited number of observations employed in the study, which restricts the model's ability to adequately train across a diverse range of observations. It is anticipated that as the dataset accumulates more data, the model's accuracy will improve, leading to better performance across the entire range of observations.

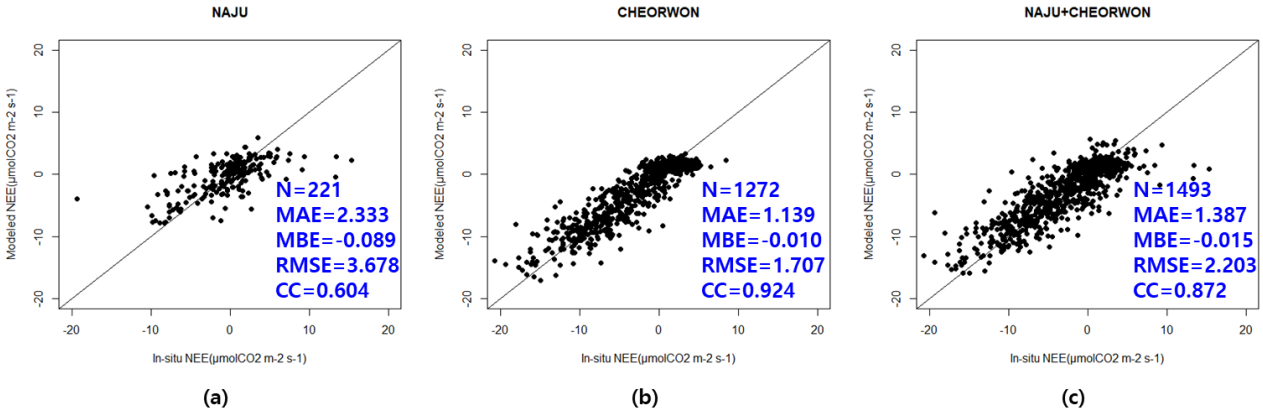

Figure 4. Scatter plots for the daily net ecosystem exchange from machine learning-based model: (a) Naju, (b) Cherowon, and (c) Naju and Cheorwon

Table 2. Accuracy statistics of the daily net ecosystem exchange from machine learning-based model

	MBE	MAE	RMSE	CC
Naju	-0.089	2.333	3.678	0.604
Cheorwon	-0.010	1.139	1.707	0.924
Naju+Cheorwon	-0.015	1.387	2.203	0.872

4. Conclusion

In this research, a machine learning-driven computational model was developed employing satellite data and meteorological information to estimate the net ecological exchange for agricultural land in Korea. Individual models were constructed for Naju and Cheorwon, locations equipped with flux towers situated in rice paddy fields. Additionally, a composite model integrating data from both regions was constructed and subjected to comparative analysis. Consequently, our findings revealed a propensity for both underestimation and overestimation in instances of very large and very small NEE observations. This phenomenon is likely attributed to the limited number of observations incorporated in the study, thereby impeding effective training across a broad spectrum of observations. In the future, with the accumulation of observations pertaining to net ecological exchange fluxes and the development of a machine learning-based calculation model leveraging high-resolution satellite and meteorological data, there is a promising prospect for achieving enhanced stability and accuracy in calculations.

Acknowledgements

This work was carried out with the support of the "Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ0162342024)" by the Rural Development Administration, Republic of Korea. This work was supported by Korea Environment Industry & Technology Institute(KEITI) through Project for developing an observation-based GHG emissions geospatial information map, funded by Korea Ministry of Environment(MOE) (RS-2023-00232066)

References

Gray J. M., S. Frolking, E. A. Kort, D. K. Ray, C.J. Kucharik, N. Ramankutty, and M.A. Friedl. 2014. “Direct Human Influence on Atmospheric CO2 Seasonality from Increased Cropland Productivity.” *Nature*, 515, 398-401. https://doi.org/10.1038/nature13957.

He, X., Zhao, K., and Chu, X. 2021. “AutoML: A survey of the state-of-the-art.” *Knowledge-Based Systems* **2021**, 212, 106622.

Jin, H., Song, Q., and Hu, X. 2019. “Auto-Keras: An efficient neural architecture search system.” *Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage*, AK, Aug. 4-8, 1946-1956.

LeDell, E. and Poirier, S. 2020. “H2O automl: Scalable automatic machine learning.” *In Proceedings of the AutoML Workshop at ICML*, Online, 17–18 July 2020; Volume 2020.

Saito, M., Miyata, A., Nagai, H., and Yamada, T. 2005. “Seasonal variation of carbon dioxide ex- change in rice paddy field in Japan.” *Agricultural for Meteorology*, 135, 93-10

Zeng N., Zhao F., Collatz G. J., Kalnay E., Salawitch R. J., West T. O., and Guanter L. 2014. “Agricultural green revolution as a driver of increasing atmospheric CO2 seasonal amplitude.” *Nature*, 515, 394–397. https://doi.org/10.1038/nature13893.