WEP Warwick Evans Publishing

International Journal of Natural Resources and Environmental Studies

ISSN: 3006-2012 (Print), ISSN: 3006-0834 (Online) | Volume 2, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijnres.v2n1.22
Journal homepage: https://wepub.org/index.php/IJNRES/index

Modeling the Ecosystem Water Use Efficiency of Croplands Using Machine Learning and MODIS Data

Xianye Meng[1, *], Xin Zheng[1], Jiaojiao Huang[1]

[1] Space Information and Big Earth Data Research Center, College of Computer Science and Technology, Qingdao University, Qingdao 266071, China

*Corresponding Author: Xianye Meng

ABSTRACT

The ecosystem water use efficiency (WUE), which is the ratio of carbon gain (gross primary productivity, GPP) to water consumption (evapotranspiration, ET), is widely regarded as a crucial link in coupling the carbon and water cycles of terrestrial ecosystems on a global scale. The estimation of cropland WUE is considerably important for improving agricultural management, simulating the carbon–water cycle processes of croplands, and enhancing crop yields. Currently, the estimate of WUE primary relies on the modeled GPP and ET data. However, differences in model structures and parameter uncertainties cause remarkable differences in the accuracy of the estimated WUE among different models. The eddy covariance technology enables the continuous observation and research of carbon and water fluxes at a site scale, ensuring simulation accuracy. However, the technology is constrained by its limited spatial coverage. In this study, we integrated data from 20 global cropland sites and built 7 machine learning models based on remote sensing to directly estimate cropland WUE, thereby expanding the spatial scope of WUE simulations, omitting the simulation process for obtaining GPP and ET, and ensuring a certain degree of precision. Results show that, compared with other machine learning algorithms, the ensemble learning model has the strongest ability to reproduce cropland WUE, with a determination coefficient of 0.90 and a root mean square error of 1.54–1.75 $gckg^{-1}H_2O$. These results indicate that a combination of machine learning methods and remote sensing can reasonably simulate the WUE of agricultural ecosystems.

KEYWORDS

Water Use Efficiency; Machine Learning; Eddy Covariance Technology.

1. INTRODUCTION

The growing population have placed unprecedented demands on agriculture and natural resources [1]. At present, approximately 1 billion people suffer from chronic malnutrition, and our agricultural system has led to the degradation of land, water, biodiversity, and climate on a global scale, resulting in stagnant or declining crop yields in certain regions [2]. To meet the future demands of global food security and sustainability, food production must increase considerably [3,4]. The global per capita share of freshwater resources has decreased by >20% over the past two decades. As the frequency of global climate change and abnormal precipitation increases, the decline in per capita freshwater resources seriously threatens the sustainable development of agriculture. Therefore, improving water use efficiency (WUE) and ensuring the full utilization of water resources is important. This approach is crucial for achieving a steady growth in crop yields [5]. For implementing this approach, the WUE of cropland must be optimized and effects of various environmental factors, such as temperature, wind speed, and incident radiation, on WUE must be explored. Analyzing the cropland WUE through climate change and studying the impact of environmental changes on agriculture are greatly

significant in preventing natural disasters and ensuring global food production. The eddy correlation method or ecological model driven by remote sensing is used to study the WUE of the ecosystem, the acquisition of WUE is first calculated using the GPP and ET of the ecosystem. Subsequently, WUE is estimated, which can be modeled directly. Machine learning (ML), a hotspot of artificial intelligence, has various applications and can be used as an effective means for modeling. Cropland is an important vegetation type in our ecosystem; hence, analyzing the WUE of cropland will help us predict drought situations and ensure crop food production. The main purpose of this study is to directly model the WUE of cropland and analyze the WUE of global cropland sites utilizing ML methods.

2. MATERIALS AND METHODS

2.1. Flux-site data

FLUXNET is an international network that connects regional networks of Earth system scientists. FLUXNET scientists use eddy covariance technology to measure the carbon, water, and energy cycles between the biosphere and the atmosphere, obtaining information on ecosystem changes. The daily scale includes meteorological data of the cropland site in our study comes from FlUXNET2015, and the download path for the dataset is as follows (http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/). This dataset includes multiple vegetation types, and we have selected cropland as our research object, which includes twenty sites. Our study uses a meteorological variable dataset used that includes seven easily obtainable variables: air temperature (Ta), precipitation (P), VPD, atmosphere pressure (Pa), wind speed (WS), incoming shortwave radiation (Rg), incoming longwave radiation (RL), carbon dioxide concentration in the atmosphere (Ca). It also includes two important parameters that help us obtain the target variable: GPP and latent heat flux (LE). We use the latent heat flux (LE_F_MDS) provided by the site data to estimate the evapotranspiration (ET) of the ecosystem. The LE data obtained from the original site data after energy closure correction, is calculated using the following formula:

$$ET = LE/\lambda \times 86400 \quad (1)$$

$$\lambda = (2.501 - 0.002361 \times T) \times 10^6 \quad (2)$$

In the equation, λ (Unit: J/kg) represents the latent heat of vaporization, 86400 is the time conversion coefficient, and T (Unit: °C) is the temperature.

2.2. MODIS data

In this study, we used five remote sensing variables: normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), leaf area index (LAI), day of land surface temperature (LST-D), and night of land surface temperature (LST-N). We obtained NDVI and EVI from MODIS remote sensing data, which are crucial indicators for monitoring vegetation phenology. Researchers have widely used these indicators in global climate change response studies and analyzed them using the enhanced TIMESAT algorithm [6,7]. LAI, which quantifies the total plant area per unit land area, serves as an important parameter to estimate ground vegetation [8]. Another significant parameter for surface–atmosphere material and energy exchange is LST, which reflects the increase in surface temperature due to solar radiation absorption by the ground. We obtained remote sensing data that matched the site location using the latitude and longitude provided in Table 1. We retrieved data from the following download path: (https://modis.ornl.gov/data.html). We directly obtain NDVI and EVI

from MO(Y)D13Q1 [9,10], with a temporal resolution of 16 days and a spatial resolution of 250 m. The LAI data corresponds to MCD15A2H [11], which has an eight-day time resolution. These variables collectively represent the cover and growth of the surface vegetation. We used MOD11A1 to obtain the LST with a spatial resolution of 1 km.

2.3. Data processing

The data processing process used in our study is as follows:

1. The time scale of the remote sensing data is standardized to the daily scale using the interpolation method and combined with our conventional meteorological data.

2. Remove the outliers of GPP and LE using a box plot and calculate the target variable WUE.

3. Fill in the missing data via the linear interpolation method.

4. Remove the nongrowing season data of the site based on latitude and longitude coordinates, eliminate data corresponding to January, February, and December in the northern hemisphere, and exclude data corresponding to June, July, and August in the southern hemisphere.

5. Within the time scale of all data at each flux site, the first 70% of the data are used as the training set and last 30% are used as the test set.

6. All attribute variables in the study are averaged and normalized to ensure that the processed data conform to the standard normal distribution.

2.4. Models based on machine learning

In this study, we used seven ML models to estimate the WUE of croplands. For our model, five meteorological factors and five remote sensing data are used as inputs, and the employed ML models include least absolute shrinkage and selection operator, random forest, external gradient boosting, light gradient boosting machine, support vector regression, multilayer perceptron, and stacking. Fig 1 presents the flowchart of our research, including data selection and model processing.

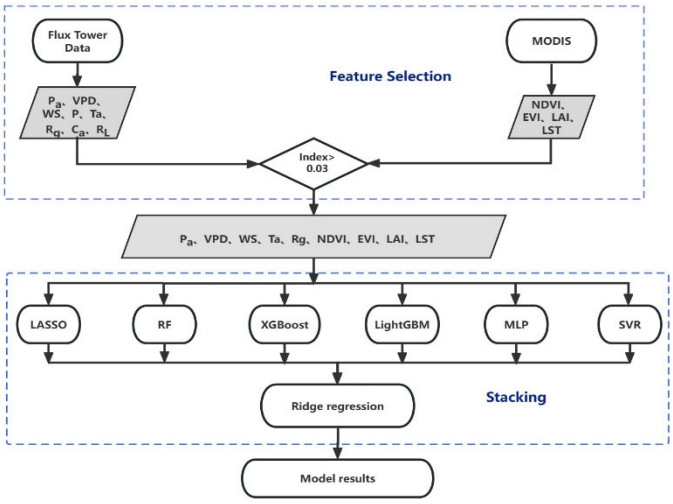

Fig ure 1. Research flowchart including feature selection and model training process.

2.5. Evaluating the performance of the models

In our study, we train seven ML models using meteorological data and remote sensing data. This approach provides a way to model WUE theoretically and directly. We divide the dataset into a

training set and a test set. We use cross-validation in model training to find the optimal parameters and use the test set to evaluate the generalization performance of the models. For model evaluation, we use two common metrics in regression models: root mean square error and coefficient of determination. In these metrics, y_i represents our true sample values, f_i represents our model predictions, and $\bar{y}$ represents the mean of the observations.

Root mean square error (RMSE): The root mean square error (MSE) measures model calculates the square of the distance (i.e., error) between the predicted and actual values to measure model merit. The smaller the value of MSE, the better the accuracy of the prediction model in describing the experimental data. The RMSE is the square root of the MSE, which can be considered as the standard deviation between the predicted and true values and is used to measure the dispersion of the predicted values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(f_i - y_i)^2}{N}} \tag{3}$$

The coefficient of determination (R^2): We use R^2 to evaluate the fitting degree of the model. It reflects the proportion of all the variance of the dependent variable that the independent variable can explain through the regression relationship. Furthermore, it indicates the proportion of the variance that the model can explain to the total variance. The value of R^2 ranges from 0 to 1, and a value closer to 1 indicates a better fit of the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{4}$$

3. RESULTS

3.1. Feature selection

To enhance the efficiency and accuracy of our model, we performed feature selection on the meteorological data in the FLUXNET dataset. This involved selecting the most influential features that show a strong correlation with the target variable. Random Forest, an integrated learning algorithm, was employed for classification and regression prediction using multiple decision trees. Random Forest has gained popularity in the fields of ML and data science due to its effectiveness and efficiency. One of the valuable functions that random forest provides is feature importance analysis, which helps in identifying features that contribute the most to the predicted results. To assess feature importance, we used the Gini importance method. This approach evaluates the significance of a feature by calculating the number of times it is used as a split node in the decision trees and degree to which it minimizes the Gini index at each node. The Gini index is a metric used to measure the impurity of a sample within a dataset. Our analysis show that out of the eight meteorological factors and remote sensing data considered, Ca, Ra, and P have the least importance in our model. Therefore, considering our goal of improving model performance, conducting more accurate analyses, and focusing on WUE-related variables, we decided to exclude these three factors from further analysis (Fig 2).

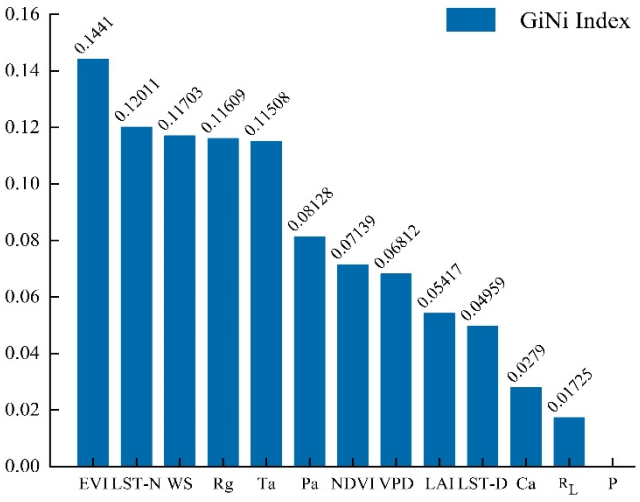

Figure 2. Gini index of input data wherein higher numbers indicate the stronger predictive power of the model.

3.2. Evaluation results of the simulated WUE

Fig 3 presents the evaluation results of the models. RF, XGBoost, LightGBM, and MLP showed better performance on the training set, with R^2 ranging from 0.89 to1. The LASSO and SVR models fit less well with R^2 of 0.12 and 0.33, respectively, after integration using the stacking model, the R^2 on the training set was 0.98 and the RMSE was 0.80 ($gCkg^{-1}H_2O$). To verify the generalization performance of the model, we tested the model on the test set and the results are shown below. The performance differences of RF, XGBoost, LightGBM, and MLP on the test set were not significant, with R^2 ranging from 0.79 to0.88 and RMSE ranging from 1.54 to 1.86 ($gCkg^{-1}H_2O$). The LASSO and SVR models fit less well on the test set, with R^2 of 0.10 and 0.33, respectively, and RMSE values of 3.78 ($gCkg^{-1}H_2\ O$) and 3.30 ($gCkg^{-1}H_2\ O$). After, using the stacking model integration, the test set achieved the optimal performance, with a R^2 value of 0.90 and an RMSE value of 1.23 ($gCkg^{-1}H_2O$).

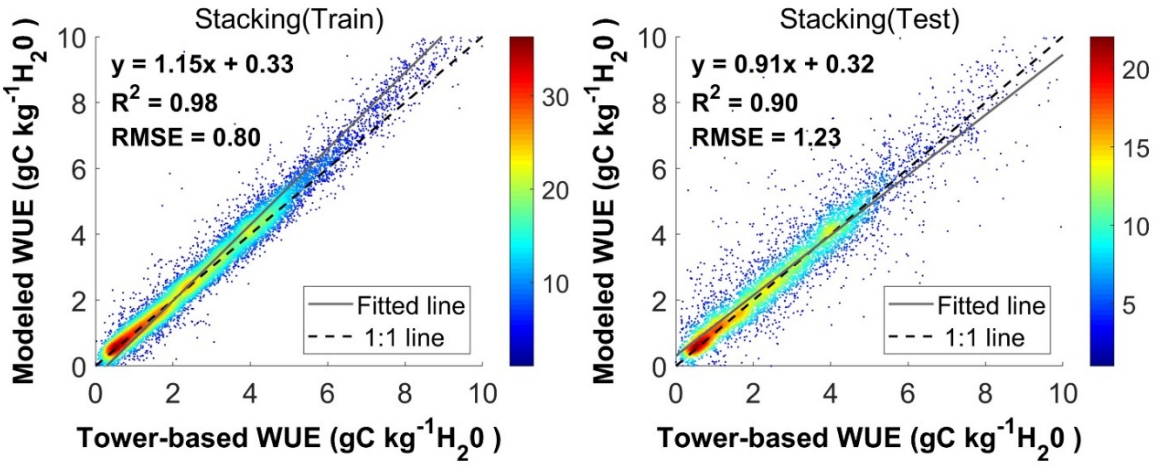

Figure 3. Evaluation of daily water use efficiency (WUE) using stacking models on the training and testing sets. The modeled WUE represents the predicted values of the model, and the tower-based WUE represents the true value of the site data. Color bar represents the data density level, red represents high data volume distribution.

4. DISCUSSION

4.1. Implications for coping with agricultural disasters

In recent years, there have been numerous natural disasters and extreme weather events. WUE, as a coupled parameter of GPP and ET, connects photosynthesis of plants with water flow between plants and nature, and it is essential for us to study changes in ecosystems to maintain their stability. ML can transcend the limitations of human cognitive data to discover patterns and capture the relationships between inputs and outputs from vast amounts of data. Therefore, our study employed five conventional meteorological variables and five remote sensing datasets as inputs of our model, and subsequently, seven ML models were trained to estimate the WUE of croplands. Herein, we propose a new method for directly obtaining WUE that surpasses our general understanding of WUE by omitting the simulation process for GPP and ET. Our research was performed on 20 cropland sites, which encompasses various crop types, and demonstrated effective simulation of WUE. We achieved high accuracy and extensive coverage for the cropland WUE by processing 10 types of input data through model training. Thus, we addressed the limited spatial range of eddy covariance technology measurements and the poor accuracy of MODIS data in WUE simulations. Benefiting from the readily available conventional meteorological data and support of MODIS remote sensing data, we could perform the high-accuracy WUE analysis in croplands without flux tower installations, obtaining WUE under specific spatiotemporal conditions. This enables the maintenance of ecosystem productivity under limited water resources, ensuring the sustainability and stability of agricultural production [12,13]. As the climatic condition changes by location and spatial scope expands, the input data for the model will differ, which allows for the WUE simulation under different geographical conditions. This facilitates the formulation of tailored water-saving strategies, improving integrated water resource management [14] and ensuring ecosystem sustainability.

4.2. Uncertainty in machine learning models

Although considerable progress has been attained in the study of the interpretability of ML models, ML is still considered a black box in many cases [15]. This is because there is a trade-off between the model interpretability and model performance, in which model performance does not depend on runtime or execution performance but rather on the accuracy of predictions [16]. In our study, the LASSO regression and SVR models exhibit high model interpretation and present low model performance. For the stacking integration model, it corresponds to a lower model interpretation and yields a higher prediction performance. For solving ML problems, we usually prefer to focus on model performance metrics such as RMSE, R^2, but metrics can only account for a part of the model prediction and performance may vary over time because of conceptual model drift caused by various environmental factors. Furthermore, the uncertainty in the input data can result in changes in the ML model performance. In the meteorological and remote sensing data used in our study, the three conventional meteorological variables Ca, Rg, and P were excluded through the importance analysis in random forest to enhance model efficiency. However, in general research on the cropland WUE, the main factors influencing model performance should be P, Ta, Ra, and extreme weather events. Upon analyzing the reasons for this difference, we suspect that this happens because of the influence of human activities, such as cropland irrigation. Owing to the existence of artificial conditions, the migration of our model will be reduced when studying other vegetation types. In addition to the meteorological data, we did not use all easily available remote sensing data for our experiments, and different data inputs will certainly have different impacts. Moreover, in the process of training our model, adjusting the model parameters is necessary to improve the generalization performance of the model in finding the optimal model. Taking random forest as an example, we need to adjust the number of trees in the forest, maximum depth of the tree, minimum number of samples required for each node before splitting, and minimum number of samples required to be included in a leaf node. Different parameter combinations can influence the performance of our model, ultimately resulting

in different model results. Additionally, for the SVR algorithm, the choice of the kernel function, penalty parameter C, and kernel parameters influence the model's performance and generalization ability. For the MLP model, the number of layers, number of neurons in each layer, type of activation function, learning rate, and other parameters directly affect the model's complexity and learning ability. The optimization algorithm may converge to local minima, which can lead to poor model performance.

5. CONCLUSIONS

An accurate assessment of the WUE of the ecosystem is of great importance in understanding the carbon–water coupling in the ecosystem and predicting the occurrence of natural disasters. Analyzing the WUE of the cropland ecosystem is highly significant in predicting droughts over croplands and ensuring grain production. In our study, we used conventional meteorological and remote sensing data to directly model WUE using seven ML methods. Further, we adjusted the important parameters of each ML model to enhance its performance. From our study, the following main conclusions can be drawn:

1. For the feature selection of WUE, we used conventional meteorological variables and remote sensing data for comprehensive analysis. Among these, WS, Ta, VPD, Rg, and Pa ranked in the top five positions in meteorological variables. This indicates that these five conventional meteorological factors highly impact the WUE analysis of our cropland sites.

2. Among the seven ML models, the integrated learning models represented by RF, XGBoost, and stacking demonstrated high performance. Moreover, the stacking model exhibited the best generalization performance on the test set and showed the best fit with R^2 reaching 0.90 and RMSE of 1.23 ($gCkg^{-1}H_2O$).

3. Using ML, we can accurately and directly analyze the WUE of cropland ecosystems. This approach skips the intermediate steps of GPP and ET simulations, simplifies the calculation process, improves calculation efficiency, and makes monitoring cropland WUE much easier.

CONFLICTS OF INTEREST

All authors declare that they have no potential competing interests.

REFERENCES

[1] Foley, J.A.; Ramankutty, N.; Brauman, K.A.; Cassidy, E.S.; Gerber, J.S.; Johnston, M.; Mueller, N.D.; O'Connell, C.; Ray, D.K.; West, P.C.J.N. Solutions for a cultivated planet. 2011, 478, 337-342.

[2] Tilman, D.; Balzer, C.; Hill, J.; Befort, B.L. Global food demand and the sustainable intensification of agriculture. Proc Natl Acad Sci U S A 2011, 108, 20260-20264, doi:10.1073/pnas.1116437108.

[3] West, P.C.; Gerber, J.S.; Engstrom, P.M.; Mueller, N.D.; Brauman, K.A.; Carlson, K.M.; Cassidy, E.S.; Johnston, M.; MacDonald, G.K.; Ray, D.K.; et al. Leverage points for improving global food security and the environment. Science 2014, 345, 325-328, doi:10.1126/science.1246067.

[4] Ray, D.K.; Mueller, N.D.; West, P.C.; Foley, J.A. Yield Trends Are Insufficient to Double Global Crop Production by 2050. PLoS One 2013, 8, e66428, doi:10.1371/journal.pone.0066428.

[5] Mueller, N.D.; Gerber, J.S.; Johnston, M.; Ray, D.K.; Ramankutty, N.; Foley, J.A. Closing yield gaps through nutrient and water management. Nature 2012, 490, 254-257, doi:10.1038/nature11420.

[6] Chang, Q.; Zhang, J.; Jiao, W.; Yao, F. A comparative analysis of the NDVIg and NDVI3g in monitoring vegetation phenology changes in the Northern Hemisphere. Geocarto International 2016, 33, 1-20, doi:10.1080/10106049.2016.1222633.

[7] Lee, R.; Yu, F.; Price, K.P.; Ellis, J.; Shi, P. Evaluating vegetation phenological patterns in Inner Mongolia using NDVI time-series analysis. International Journal of Remote Sensing 2010, 23, 2505-2512, doi:10.1080/01431160110106087.

[8] Jeganathan, C.; Dash, J.; Atkinson, P.M. Characterising the spatial pattern of phenology for the tropical vegetation of India using multi-temporal MERIS chlorophyll data. Landscape Ecology 2010, 25, 1125-1141, doi:10.1007/s10980-010-9490-1.

[9] Huete, A.; Justice, C.; Van Leeuwen, W.J.A.t.b.d. MODIS vegetation index (MOD13). 1999, 3, 295-309.

[10] Solano, R.; Didan, K.; Jacobson, A.; Huete, A.J.T.T.U.o.A. MODIS vegetation indices (MOD13) C5 user's guide. 2010.

[11] Fang, H.; Zhang, Y.; Wei, S.; Li, W.; Ye, Y.; Sun, T.; Liu, W. Validation of global moderate resolution leaf area index (LAI) products over croplands in northeastern China. Remote Sensing of Environment 2019, 233, doi:10.1016/j.rse.2019.111377.

[12] Olsen, J.; Dettinger, M.; Giovannettone, J.J.B.o.t.A.M.S. Drought Attribution Studies and Water Resources Management. 2023, 104, E435-E441.

[13] Singh, A.J.A.W.M. Simulation–optimization modeling for conjunctive water use management. 2014, 141, 23-29.

[14] Agarwal, A.; de los Angeles, M.S.; Bhatia, R.; Chéret, I.; Davila-Poblete, S.; Falkenmark, M.; Villarreal, F.G.; Jønch-Clausen, T.; Kadi, M.A.; Kindler, J. Integrated water resources management; Global water partnership Stockholm: 2000.

[15] Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016; pp. 1135-1144.

[16] Doshi-Velez, F.; Kim, B.J.a.p.a. Towards a rigorous science of interpretable machine learning. 2017.