Ecological Indicators 83 (2017) 41–52

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Fuzzy Rough Set algorithm with Binary Shuffled Frog-Leaping (BSFL-FRSA): An innovative approach for identifying main drivers of carbon exchange in temperate deciduous forests

Yueju Xue[a], Yueming Chen[b,c,d], Yueming Hu[b,c,d,*], Hanming Chen[a]

[a] Key Laboratory of Information Acquisition and Application in Agriculture, Guangzhou Science Technology and Innovation Commission, Guangzhou 510640, China
[b] Guangdong Province Key Laboratory of Land use and consolidation, Guangzhou 510640, China
[c] Key Laboratory of Construction Land Transformation, Ministry of Land and Resources, South China Agricultural University, Guangzhou 510642, China
[d] College of Natural Resources and Environment, South China Agricultural University, Guangzhou 510642, China

ARTICLE INFO

Keywords:
Carbon flux
Variables selection
Fuzzy rough set
America deciduous forests
Driving variable contribution rate
Binary shuffled frog-leaping

ABSTRACT

The spatial variability, driving forces, and uncertainties of the net ecosystem exchange (NEE) of carbon between the temperate forests and the atmosphere remain elusive. Here, we proposed a fuzzy rough set algorithm with binary shuffled frog leaping (BSFL-FRSA) to identify main driving variables and define the contribution rate of main drivers to NEE. As a case study, we applied the approach to nine deciduous forest eddy covariance flux sites in the northeastern United States. The results show that the BSFL-FRSA effectively retained most information using just a few variables, and it performed better than the GA-FRSA (fuzzy rough set with genetic algorithm). Temperature, radiation, and soil water content were identified as the most influential variables (impact in descending order) to NEE across all sites. Soil temperature was the most important variable explained 59.6% of the NEE variance. Soil temperature and net radiation together, explained 72.7% of the NEE variance, was the most important two variables among all possible two-variable combinations. The most influential three variables on NEE among all possible three-variable combinations were soil temperature, net radiation, and soil water content or relative humidity (explained 81.1% of the NEE variance). The variance attribution approach presented here is generic and can be applied to other studies; the dominant influence of soil temperature begs for accurate characterization of soil temperature dynamics in time and space particularly in the global warming context.

1. Introduction

Rising concentration of the atmospheric CO_2 causes a serious threat to the climate system and sustainable development of human societies (Chen et al., 2011a, 2011b; Xiao et al., 2011; IPCC, 2013). Pan et al. (2011) reported that temperate forests contributed about 31% to the global terrestrial carbon sink between 1990 and 2007 and play an essential role in mitigating global warming. Temperate deciduous forests are an important type of forests in the North America. And identifying the meteorological and environmental controls of NEE over these ecosystems can help predict the potential impacts of climate change scenarios on terrestrial ecosystem carbon cycles and their feedback on the global climate change (Pan et al., 2011; Froelich et al., 2015; Tan et al., 2015). However, the contribution of either individual variables or their combinations to NEE is not well understood due to complex interactions among variables and uncertainty of the carbon cycle in forest ecosystems.

The global carbon flux observation network (http://www.fluxnet.ornl.gov) provides high-quality data support for investigating how meteorological and environmental variables drive variations in half-hourly forest NEE (Baldocchi and Wilson, 2001; Hollinger et al., 2004; Liu et al., 2004). Earlier studies have analyzed the driving variables of NEE in temperate deciduous forests (e.g., Canadell et al., 2000; Fan et al., 1998; Tang et al., 2012; Ishtiaq and Abdul-Aziz, 2015) with efforts to reveal the relationships between single specific meteorological (or environmental) variables and NEE on the daily, seasonal, or annual scale (Bergeron et al., 2007; Curtis et al., 2002; van Dijk et al., 2005; Schmidt et al., 2011; Chen et al., 2011a, 2011b; Wu and Chen, 2013; Yuan et al., 2007). For example, photosynthetically active radiation, soil temperature, air temperature and precipitation were observed to influence CO_2 uptake in a mixed deciduous and coniferous forest on seasonal or monthly time-scales (Froelich et al., 2015). However, none

* Corresponding author at: Present address: Guangdong Provincial Key Laboratory of Land Use and Consolidation, South China Agricultural University, Guangzhou 510642, China.
E-mail address: yueminghugis@163.com (Y. Hu).

http://dx.doi.org/10.1016/j.ecolind.2017.07.039
Received 4 January 2017; Received in revised form 27 June 2017; Accepted 16 July 2017
Available online 30 July 2017

of these studies have applied models to clarify and quantify the most important driving variable that can, to a greater extent explain the variance in NEE at a half-hourly time scale.

Process-oriented biosphere models and data-driven analytic approaches are used to simulate or predict carbon fluxes (Liu et al., 2008; Wu et al., 2014b). Many process-oriented biosphere models have been developed including Dynamic Global Vegetation Models (DGVMs) (Müller and Lucht, 2007), Interannual Flux Tower Upscaling Experiment (IFUSE) (Desai, 2010), ECOSYS (Grant et al., 2012), DAYCENT (Parton et al., 1998), BIOME-BGC (Running and Hunt, 1993), Erosion Deposition Carbon Model (EDCM) (Liu et al., 2003; Tan et al., 2009; Wu et al., 2014a), Empirical Mode (Shao et al., 2014), and other terrestrial biogeochemical models (TBMs) (Fang and Michalak, 2015). These models use air temperature, precipitation, vapor pressure deficit, soil temperature, soil water content, or net radiation as inputs (Chen et al., 2011a, 2011b; Shao et al., 2014; Sims et al., 2008) to elucidate and quantify daily, weekly, monthly, or annual variations in NEE. In general, model predictions appear to be quite uncertain because of uncertainties in various input datasets, model parameters, and complex and inadequate model structure (Chen et al., 2011a, 2011b; Liu et al., 2011; Schwalm et al., 2015). Therefore it is still a major challenge to establish advanced accurate mechanistic models to account for the variability and uncertainty of the terrestrial ecosystem carbon cycle processes across space and time.

Data-driven analytic approaches play important roles in understanding the underlying processes and building desirable process-based models (Gupta and Nearing, 2014). We may gain new understanding about the characteristic behaviors and controlling variables of ecosystem processes by applying innovative data mining techniques and then to strengthen relevant theories and models (Gupta et al., 2010). For example, Stoy et al. (2009) used an innovative orthonormal wavelet transformation to evaluate the variability in NEE from 253 sites at multiple time scales and proved a homeostatic mechanism by which ecosystem CO_2 uptake dampens instantaneous low frequency meteorological variability (Odum, 1969). Schmidt et al. (2011) presented a self-organizing feature map neural network approach to successfully assess the influence of various environmental parameters on the carbon fluxes over vegetated areas. Mueller et al. (2010) presented a geostatistical regression approach with the Bayes Information Criterion to identify an optimal set of environmental variables that sufficiently explain the observed variability in NEE.

The data-driven approaches for identifying main drivers of carbon flux can be roughly grouped into two categories: filter and wrapper. The filter approaches select the feature subsets independent of the classification/prediction model. Examples include using Pearson correlation analysis to analyze the linear correlation between variables and NEE in eight temperate forests in North America (Ishtiaq and AbdulAziz, 2015) and using complex regression tree models to select the main variables for carbon flux prediction and then quantifying the relative frequency of use and importance of variables for prediction from the complex regression tree models (Wylie et al., 2007). In the wrapper model, the prediction accuracy of a predetermined learning algorithm is used to evaluate the feature subsets. Typical methods include random forests (e.g., Woodall et al., 2015, identify the variables that explain the most variation in total carbon flux), and genetic algorithm (GA) and neural network (NN) (e.g., Xue et al., 2006, driving factors selection). These wrapper approaches generally outperform the filter approaches in the final predictive accuracy of a learning machine. However, they are more computationally expensive than the filter approaches, and exhibit relatively poor generalization. The generalization ability of variables selection for explaining and predicting NEE is a very important issue. Moreover, the existing studies on quantifying the relationship between carbon flux and meteorological and environmental variables do not explicitly analyze the influence of combinatorial variables on NEE. Because of the presence of abundant uncertainty and the interference of irrelevant or misleading noisy variables, the ability to handle imprecise and inconsistent information on carbon flux issue has become one of the most important requirements for process understanding and selection of effective variables for process representation (Zhang et al., 2012).

Fuzzy Rough Sets Algorithm (FRSA) (Degang and Suyun, 2010) can avoid the adverse reduction effect caused by improper discretization and retain the original data information better than Rough Sets, providing an effective method for reducing variable uncertainty of data sets (Jensen and Shen, 2005). However, the minimum variable (attribute) reduction has been proven to be a non-deterministic polynomial-time hard (NP-hard) problem (Jensen and Shen, 2005), and classic fuzzy rough intensive Jane algorithm exists with such problems as complex search space and low reduction efficiency. In order to obtain the dataset with the minimum reduction in the number of attributes effectively, the rough sets reduction algorithm coupled with genetic algorithm and the ant colony algorithm were introduced and validated (Saha et al., 2013), but the relevant literature in this area is still very limited. The Shuffled Frog-Leaping Algorithm (SFLA) is a heuristic sub population-based cooperative search algorithm (Eusuff and Lansey, 2003) and has been successfully used to optimize water resource distribution networks, parametric of analog integrated circuits and job shop scheduling problems. By combining the advantages of genetic algorithm for the model based on particle swarm algorithm and group behavior, SFLA, taking the Shuffled Complex Evolution as an executive search framework, has simple models, better overall search ability and astringency (Eusuff et al., 2006).

Here we proposed an algorithm that integrated the fuzzy rough set (Degang and Suyun, 2010) with the binary shuffled frog-leaping algorithm (Gomez-Gonzalez et al., 2013) to identify the main variables driving the NEE across the US deciduous forests without considering either detailed physiological or site specific information. After building feasible algorithms (or models) to select the fewest number of variables that dominate specific NEE, we defined the driving variable contribution rate to quantify the influence of the meteorological or environmental variables on NEE.

2. Materials and methods

2.1. Sites

Eddy flux estimates of the exchange of CO_2, water vapor, and energy were derived from the covariance of high frequency fluctuation in vertical wind velocity and CO_2 concentration (Baldocchi et al., 1988). Nine flux tower sites of deciduous forests were included in this study. These sites are located in the North-America in a latitudinal range from 38.74°N to 46.73°N, and a longitude range from 71.29°W to 92.20°W (Fig. 1). According to the Köppen classification (Peel et al., 2007), the climate types range from Humid Mesothermal Climates to Humid Microthermal Climates (classes C and D) (see Table 1).

The original, not the gap-filled, half-hourly data from the LaThuile FLUXNET dataset (http://www.fluxdata.org) were used in this study to minimize imported errors potentially derived from the gap-filling procedures. In order to make the used data maximally explain how the variables influence NEE of temperate deciduous forests, all variables that potentially influence carbon fluxes should be retained. Therefore, we did not include deciduous forest sites with incomplete data, and only nine deciduous broadleaved forest sites were included.

Data missing in carbon flux observations can directly affect the variable selection and carbon flux prediction. Therefore, a three-step procedure was used to exclude outliers in the data. Firstly, observations during rainy periods were removed, because eddy-covariance measurements become unreliable during the rainy period (Baldocchi et al., 2000; Lee, 1998). Because precipitation data were not available at a half-hourly scale, we used soil moisture content and relative humidity to represent precipitation as suggested by Ishtiaq and Abdul-Aziz (2015), even though precipitation has a main effect on the interannual NEE (Piao et al., 2013; Schmidt et al., 2011). Secondly, the nighttime