Modeling of the atmospheric CO_2 concentration using Random Forest Model

Zhaleh Siabi[1],Samereh Falahatkar*[2] and Seyed Jalil Alavi[3]

[1,2,3] Faculty of Natural Resources and Marine Sciences, Tarbiat Modares University, Iran

Email: z.siabi@gmail.com

Email: samereh.falahatkar@modares.ac.ir

Email: j.alavi@modares.ac.ir

Key words: IRAN, OCO-2, Random Forest, Remote sensing, Climate change

Abstract

In this research, we were modeling the relationship among of atmospheric CO_2 concentration with environmental variables in April and August of 2015 in Iran. For this goal, we used CO_2 concentration data gathered from OCO-2 satellite and LST, NDVI, LAI and NPP data collected from MODIS products. Temperature, wind direction and wind speed data were downloaded from ECMWF database. We utilized three layers of Organic carbon stocks data in two available depths of 5-15 cm and 15-30 cm which comes from Soil Grids1km - Global Soil Information and mean of two layers were resulting a layer with depth of 5-30 cm. We also extracted land cover information from Google Earth and fossil fuel data of the N.I.O.P.D.C for whole of Iran. Random Forest model was applied for this research. To probe the model validation, we used the cross validation for April and August that resulted R^2=0.61, RMSE=1.34 and R^2=0.66, RMSE=1.14, respectively. Among 16 variables used at this model in April, four variables of temperature, wind speed, LST and wind direction were recognized as prominent parameters, where OCS_5_30 has been recognized as less important one. In August, those prominent parameters, following the order of importance, were temperature, wind speed, LST and NPP where the result revealed that the land cover were less important parameter. The results of this research revealed that the most effective and important variables in atmospheric CO_2 concentration are temperature and wind speed in Iran which is useful for sustainable management.

1. Introduction

Carbon dioxide has the most important role in global warming and climate change. Therefore, the scientific and effective control of CO_2 concentration in the atmosphere has vital importance in aspect of management at large scale. Therefor a deep understanding about the important sources of emission and absorption of this gas is necessary. In order to obtain this object, the modeling of relationship between CO_2 concentration and environmental variables which affect the emission and absorption of CO_2 will be valuable. There are two main ways for measuring and monitoring of CO_2 concentration in atmosphere including the ground-based sites and satellite-based monitoring. The ground-based stations measure CO_2 concentration directly at a near land surface. However, these stations have some limitations such as gaps in space and coverage (Li et al, 2016). In contrast, the satellite's data has a primary benefit of well spatial coverage, low cost and simple accessibility (Yue et al, 2016). The important sensors that measure the CO_2 concentration including SCIAMACHY, GOSAT and OCO-2. The OCO-2 satellite was launched in 2nd July 2014 by National Aeronautics and Space Administration (NASA) (Frankenberg et all, 2014). The OCO-2 is a sun-synchronous satellite (NASA, 2016) and captures data of CO_2 concentrations from the elevation of 705 km above the land surface. The OCO-2 satellite measures and monitors CO_2 concentrations in the local and continental scales with high spatial resolution (3km^2) in near infrared wave length range (Frankenberg et all, 2014). To reach the best measurement of CO_2 concentrations, it has a narrow spectrum range consisting of 0.76, 1.6 and 2 Micrometer and measurement accuracy equal or less than 1 ppm (Frankenberg et al, 2014). Literature reviews show that the study of carbon cycle between the atmosphere and land was done with high accuracy when integrated the modeling and remotely sensed data and field measurement. (Guo et al, 2012). For example, Guo et al. (2012) used the GOSAT satellite data and MODerate resolution Imaging Spectroradiometer (MODIS) products

data, including temperature, vegetation and productivity, from June 2009 to November 2011 and launched a model called the TVP model for the five continents Africa, Australia, Eurasia, North America and South America. Shim et al. (2013) studied the CO_2 variations in the East Asian region. They found that CO_2 concentration was highest in the spring and in summer, due to increased leaf density and high levels of photosynthesis. Guo et al (2013) investigated the spatial distribution of greenhouse gas in arid and semi-arid regions of East Asia using Kriging method. They found that the frequency of CO_2 gas was primarily influenced by plant photosynthesis, soil respiration and evapotranspiration.

In accordance with the specific climate conditions of Iran such as arid and semi-arid regions, the large temperature range, and rainfall shortage, the effect of climate changes will be severe for it. As CO_2 is the main factor in global warming and climate change, it is necessary to monitor and manage the concentration of this gas by an effective method. Our object is the modeling of the relationship between CO_2 concentration and environmental variables over Iran in order to understand the importance of each of the environmental variables for carbon dioxide management.

2. Material and Methods

2.1 study area

The study area is located in Iran, Middle Eastern part of Asia; it extends between 25°-40° N latitudes and 44°- 64° E longitudes. The location of the study area is shown in Fig1. The mean elevation of Iran is 1200 m above sea level and its mean annual precipitation and temperature are 246 mm and 18.2^{o} C, respectively.

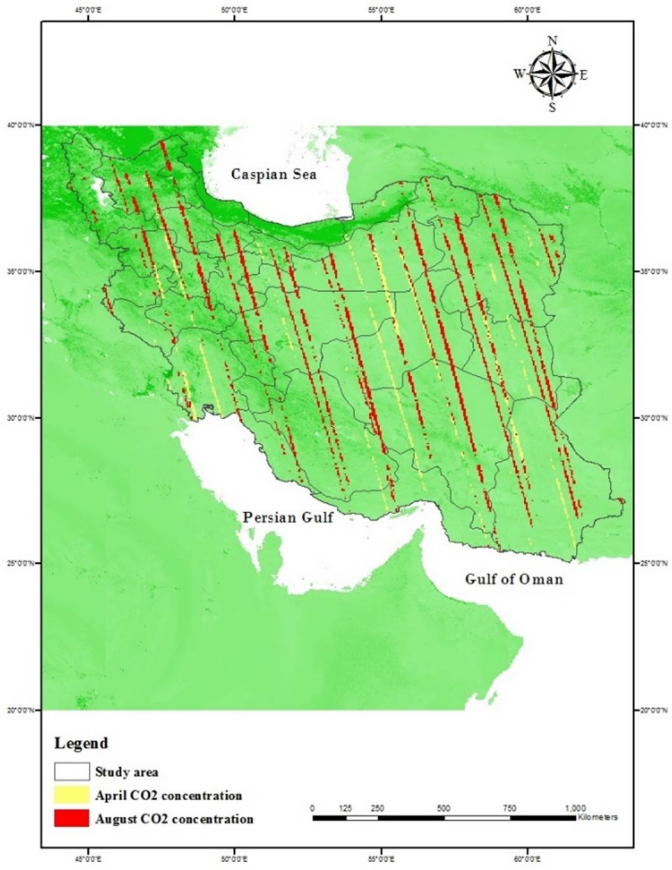

Figure 1- the study area

2.2 used data

In this study, for modeling of CO_2 concentration and environmental variables, the OCO-2 level 2 data in April and August, 2015 was used. In order to obtain the monthly CO_2 concentration, we were used the fishnets with size of 0.05×0.05 degrees in whole of study area and then calculated the average of daily data in each month. The environmental variables in this research including Normalized Difference Vegetation Index (NDVI), Net Primary Productivity (NPP), Leaf Area Index (LAI) and Land Surface Temperature (LST) from MODIS products. Furthermore, we used the meteorological parameters such as air temperature in 2 meters above the surface, 10 meters high wind speed and wind direction which obtained from European Centre for Medium-range Weather Forecasts (ECMWF) predicted data. The wind direction data have been classified into the following four categories: category 1 is 0-45 and 315-360 degrees for north direction, category 2 is 45-135 degrees for east direction, category 3 is 135-225 degrees for south direction, and category 4 is 225-315 degrees for west direction. We have utilized organic carbon stocks data in two available depths of 5-15 cm and 15-30 cm which are obtained from Soil Grids1km - Global Soil Information. Characteristics of the used data including the spatial and temporal resolution are presented in Table 1. We have extracted the land cover class from Google Earth pro (v7.1.5.1557) in nine categories (Table 2) in order to survey the land cover information for each box of fishnet. For more precise modeling, annual average of fossil fuel consumption data (Gasoline, Kerosene, Gasoil and Mazut) was used. This information was collected from the National Iranian Oil Products Distribution Company for each province of Iran in 2015. We studied the effect of the total fossil fuel consumption in the emission of CO_2 by summing the data of the mentioned above. As shown in Table 1, environmental variables are different regarding spatial and temporal resolutions. Therefore, in this study, the time resolution of all data was converted into monthly basis. In the case of data with shorter time resolution (e.g. 4 days and 16 days), it was done through averaging. On the other hand, to reconstruct the spatial resolution of the different variables, the resampling method was used and all the layers of information were converted to 250*250 meters.

Tabel 1. The MODIS products and meteorological variables characteristics

Data	product name	Spatial resolution	Temporal resolution
NDVI	MOD13Q1	250 m	16 days
NPP	MOD17A2-M	0.1°	Monthly
LAI	MCD15A3H	500 m	4 days
LST	MOD11C3	0.05°	Monthly
T-2m	2m temperature	15 km	Monthly
wind_speed	10m wind speed	15 km	Monthly
Wind_direction	10m wind direction	15 km	Monthly
OCS (in different depth)	Soil Grid 250m	250m	yearly

Table 2. Land Cover categories definition

Land cover category	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Land Cover Type	Forest	Rangeland	Agriculture	Garden	Bult	Water body	Bare	Tree pastures	Mixed area

2.3 Random Forest

After preparation of data, as described in the 2.2 section, the Random Forest Model has been applied using R software (v.3.4.0). The Random Forest model was introduced in 2001 by Breiman (Genuer et al, 2015). This model is one of the tree-based models and consists of many decision trees (including classification or regression trees) (Breiman et al, 2001). The training of each decision tree in a random forest is done by random selection of Bootstrap samples consisting of about two-thirds of the main dataset components. For each Bootstrap instance, a regression or classification tree is grown, and the final prediction of the model is made by collecting predictions from each tree. The process of collecting results in the regression random forest method is performed using the mean of all predictions, while in the classification random forest, it is performed by taking into account the highest score for a class. The main components of a random forest include the number of random variables for splitting each node and the number of trees in the forest (Liaw and Wiener, 2002). The number of random variables can be either determined by trial and error or set by the default value $p/3$ (where p is the number of independent variables considered in the model). In this research, the performance estimation of the random forest model was carried out by cross validation method with R^2 and RMSE indices.

3. Results

To evaluate the relation between the CO_2 concentration and the environmental variables, we have run a Random Forest model as described in 2.3 section for April and August months in 2015. The significance of each environmental variable in estimation of carbon dioxide values is determined based on the amount of error increasement if the variable is eliminated from the modeling process. The results of Random Forest modeling of CO_2 concentration in terms of the environmental variables is illustrated in Figure 2 and 3. According to the results (figures 2 and 3), in April the most effective variables in the estimation of CO_2 concentrations are air temperature, wind speed, LST and wind direction, respectively. In Aguest, the most effective variables are air temperature, wind speed, LST and NPP, respectively. It is worth to mention that for both of months land cover and organic carbon stock (5-30 cm layer) have a little effect on CO_2 concentration in our findings.

The accuracy assessment of the results of Random Forest modeling is illustrated in Table 3. According to results of assessment R^2 = 0.61, RMSE = 1.34 for April and R^2 = 0.66, RMSE = 1.14 for August was observed, respectively. The accuracy assessment shows a completely acceptable performance of Random Forest modeling in predicting the CO_2 concentration using the selected variables.

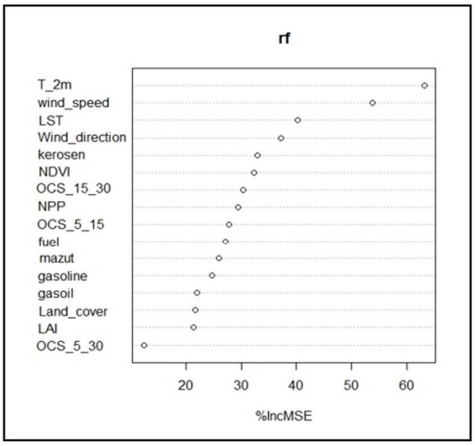

Figure 2. the April variables importance

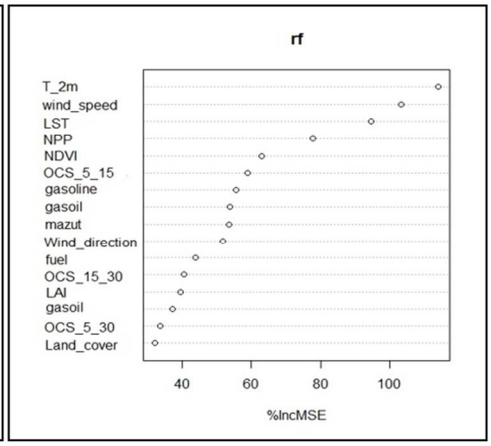

Figure 3. the August variables importance

Table 3. results of model accuracy assessment

Month	R^2	RMSE
April	0.61	1.34
August	0.66	1.14

4. Discussion

According to the modeling results that described in the results sections, air temperature and LST are the most important variables affecting the spatial distribution of carbon dioxide gas in Iran in months of April and August. This result is consistent with the results of Mousavi et al in 2016 which showed a positive and significant relationship between temperature and CO_2 concentration in Iran using GOSAT data. In addition, according to the results, it can be concluded that due to the presence of most of Iran's ecosystems in arid and semi-arid regions, land cover parameters such as NDVI, NPP and LAI are less important than climate variables. The same observation is reported by Guo et al. who carried out similar studies for East Asia in 2013 and concluded that temperature is the key parameter in CO_2 concentration in arid and semi-arid regions.

The results of our study indicate that the average concentration of CO_2 in Iran decreased from 401.61 ppm in April to 397.81 ppm in August. In the other hand, the importance of vegetation variables such as NDVI and NPP is increased from April to August. This is because in the onset of the spring season (i.e. April), due to the gradual increase in the temperature of the air, respiration of soil microorganisms increases while vegetation growth does not reach its maximum. Therefor concentration of CO_2 gas is higher in the spring than summer (Mousavi et al, 2016). In the other hand, vegetation cover is at peak in summer than spring over Iran and mean of CO_2 concentration decrease by vegetation as a natural sink. Note that respiration and photosynthesis are the most important factors in the exchange of carbon between the atmosphere and the biosphere on a large scale, and the change in each of these factors will play an important role in the concentration of carbon dioxide (Sun et al., 2001).

Wind is important factor in the diffusion, dilution and dispersion of gases in the atmosphere (Britter et al., 1989 and Oldenburg et al., 2004). As the wind velocity increases to 2 m/s, it can sweep and dilute the carbon dioxide gas of the surface (Chow et al., 2009). The results of our study agrees with the mentioned facts since wind speed was the second important variable affecting carbon dioxide emissions in both months of April and August in our study. Unfortunately, the data of fossil fuel are available yearly for whole of Iran, therefore their effect have not been shown logically in this modeling.

5. references

Breiman, L., 2001. Random forests. journal of Machine learning, 45(1), pp. 5-32.

Britter, R. E. (1989). Atmospheric dispersion of dense gases. Annual review of fluid mechanics, 21(1), pp. 317-344.

Chow, F. K., Granvold, P. W., and Oldenburg, C. M., (2009) "Modeling the effects of topography and wind on atmospheric dispersion of CO_2 surface leakage at geologic carbon sequestration sites", Energy Procedia, 1(1): 1925-1932.

Frankenberg, C., O'Dell, C., Berry, J., Guanter, L., Joiner, J., Kohler, Ph., Pollock, R., E.Taylor, T., 2014. Prospects for chlorophyll fluorescence remote sensing from the Orbiting Carbon Observatory-2. Remot Sensing of environment, 147, pp 1-12.

Genuer, R., Poggi, J.M., Tuleau-Malot, Ch., Villa-Vialaneix, N., 2015. Random forests and big data.

Guo, M., Wang, X., Li, J., Yi, K., Zhong, G., Tani, H., 2012. Assessment of global carbon dioxide concentration using MODIS and GOSAT data. Sensors, 12(12), pp. 16368-16389.

Guo, M., Wang, X.F., Li, J., Yi, K.P., Zhong, G. S., Wang, H. M., Tani, H., 2013. Spatial distribution of greenhouse gas concentrations in arid and semi-arid regions: A case study in East Asia. Journal of Arid Environments, 91, pp. 119-128.

http://www.nasa.gov/mission_pages/oco2/overview., page last updated:July 31, 2015-last visited at 07/05/2016.

IPCC, 2007. Climate change- synthesis report. Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Rome.

Li, Y., Deng, J., Mu, C., Xing, Z., Du, K., 2014. Vertical distribution of CO2 in the atmospheric boundary layer: Characteristics and impact of meteorological variables. Atmos Environ, 91, pp. 110–117.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News, (2/3), pp. 18-22.

Mousavi, S, M., Falahatkar, S., Frajzadeh, M., 2016. Monitoring of Spatial and Temporal Distribution of CO_2 and CH_4 Greenhouse Gases using GOSAT Satellite Data in IRAN. In Partial Fulfillment of the Requirement for the Degree of Master of Sciences (MS.c). Tarbiat Modarres University, Faculty of Natural Resources and Marine Sciences, pp. 1-108.

Oldenburg, C. M., and Unger, A. J., (2004) "Coupled vadose zone and atmospheric surface-layer transport of carbon dioxide from geologic carbon sequestration sites", Vadose Zone Journal, 3(3), pp. 848-857.

Sun, Y., 2001. The study on CO_2 of Karst Eco-system of vertical zone in Jinfo Mountain in summer. PhD thesis.

Shim, Ch., Lee, J., Wang, Y., 2013. Effect of continental sources and sinks on the seasonal and latitudinal gradient of atmospheric carbon dioxide over East Asia. Atmospheric Environment, 79, pp. 853-860.

Yue, T, X., Zhang, L, L., Zhao, M, W., Wang, Y, F., Wilson, J., 2016. Space- and ground-based CO2 measurements: A review. Science China Earth Sciences, 59, pp. 2089–2097.